

USING MATHEMATICS CURRICULUM BASED MEASUREMENT
AS AN INDICATOR OF STUDENT PERFORMANCE ON STATE STANDARDS

A Dissertation

by

LINDA DE ZELL HALL

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2009

Major Subject: Educational Psychology

USING MATHEMATICS CURRICULUM BASED MEASUREMENT
AS AN INDICATOR OF STUDENT PERFORMANCE ON STATE STANDARDS

A Dissertation

by

LINDA DE ZELL HALL

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

| | |
|---------------------|----------------|
| Chair of Committee, | Richard Parker |
| Committee Members, | Linda Parrish |
| | Luana Zellner |
| | Mack Burke |
| Head of Department, | Victor Willson |

December 2009

Educational Psychology

ABSTRACT

Using Mathematics Curriculum Based Measurement
as an Indicator of Student Performance on State Standards. (December 2009)

Linda De Zell Hall, B.B.A., Texas Tech University;

M.Ed., Texas A&M University

Chair of Advisory Committee: Dr. Richard Parker

Math skills are essential to daily life, impacting a person's ability to function at home, work, and in the community. Although reading has been the focus in recent years, many students struggle in math. The inability to master math calculation and problem solving has contributed to the rising incidence of student failure, referrals for special education evaluations, and dropout rates. Studies have shown that curriculum based measurement (CBM) is a well-established tool for formative assessment, and could potentially be used for other purposes such as a prediction of state standards test scores, however to date there are limited validity studies between mathematics CBM and standard-based assessment. This research examined a brief assessment that reported to be aligned to national curriculum standards in order to predict student performance on state standards-based mathematics curriculum, identify students at-risk of failure, and plan instruction. Evidence was gathered on the *System to Enhance Educational Performance Grade 3 Focal Mathematics Assessment Instrument* (STEEP3M) as a formative, universal screener. Using a sample of 337 students and 22 instructional staff, four qualities of the STEEP3M were examined: a) internal consistency and criterion related validity (concurrent); b) screening students for a multi-tiered decision-making process; c) utility for instructional planning and intervention recommendations; and d) efficiency of administration, scoring, and reporting results which were the basis of the four research questions for this study. Several optimized solutions were generated from Receiver Operator Curve (ROC) statistical analysis; however none demonstrated that the STEEP3M maximized either sensitivity or specificity. In semi-structured interviews

teachers reported that they would consider using the STEEP3M, however only as a part of a decision-making rubric along with other measures. Further, teachers indicated that lessons are developed before the school year starts, more in response to the sequence of the state standards than to students' needs. While the STEEP3M was sufficiently long enough for high-stakes or criterion-referenced decisions, this study found that the test does not provide sufficient diagnostic information for multi-tiered decision-making for intervention or instructional planning. Although practical and efficient to administer, the conclusions of this study show the test does not provide sufficient information on the content domain and does not accurately classify students in need of assistance.

DEDICATION

This dissertation is dedicated to my family who encouraged and supported me through this process

ACKNOWLEDGEMENTS

“Alone we can do so little; together we can do so much.”

Helen Keller

It is a privilege to live in the United States where our educational system is based upon the belief that all children can learn and have the right to access a free public education. My gratitude goes to the developers of the curriculum based measurement who shared the instrument for use in this study, and to the educators who participated in this research. I would like to express my appreciation to my doctoral committee for sharing their guidance, expertise, and time throughout this process. Finally, and most importantly, thanks to my family for being my everything and helping me achieve my goals.

TABLE OF CONTENTS

| | Page |
|--|------|
| ABSTRACT | iii |
| DEDICATION | v |
| ACKNOWLEDGEMENTS | vi |
| TABLE OF CONTENTS | vii |
| LIST OF FIGURES..... | xiii |
| LIST OF TABLES | xiv |
| CHAPTER | |
| I INTRODUCTION..... | 1 |
| Mathematics is Fundamental..... | 1 |
| Employment Outlook | 2 |
| Problem Statement | 3 |
| Research on Mathematics..... | 3 |
| Mathematics Curriculum Based Measurement | 5 |
| Data Collection Methodology | 5 |
| Purpose of Study | 6 |
| System to Enhance Educational Performance (STEEP) | 6 |
| Research Questions | 7 |
| Organization of Study | 7 |
| II REVIEW OF LITERATURE..... | 8 |
| Legacy of A Nation at Risk..... | 8 |
| Standards-based Assessment..... | 9 |
| Assessment for Early Identification | 10 |
| Research on CBM | 10 |
| Economic Factors | 11 |
| Policy Reforms | 11 |
| Students with Disabilities under IDEA 2004 | 12 |

| CHAPTER | | Page |
|---------|--|------|
| | National Reading Panel..... | 13 |
| | National Mathematics Advisory Panel | 14 |
| | Initial Feedback on the NMAP Report..... | 15 |
| | Additional Information on Mathematics Research | 15 |
| | Defining Mathematics | 16 |
| | Multi-tiered Support Systems for Students | 16 |
| | Criteria for Useful Assessment Procedures..... | 17 |
| | Curriculum Based Assessments | 17 |
| | Models of CBA | 17 |
| | Differentiated Features of CBM..... | 18 |
| | Formative Assessment..... | 18 |
| | CBM as Formative Assessment | 19 |
| | Summative Assessment..... | 19 |
| | Standards-based Curriculum | 19 |
| | National Council of Teachers of Mathematics (NCTM)..... | 20 |
| | NCTM Focal Points | 20 |
| | Curriculum Focal Points for Grade 3 | 21 |
| | Texas Essential Knowledge and Skills..... | 21 |
| | Screening of Student Performance | 22 |
| | Criterion-referenced Tests..... | 22 |
| | Standards-referenced Tests | 23 |
| | Statewide Assessment in Texas..... | 23 |
| | Texas Assessment of Knowledge and Skills..... | 24 |
| | CBM and State Standards | 25 |
| | System to Enhance Educational Performance (STEEP) | 25 |
| | Identifying Mathematical Problems | 25 |
| III | METHOD..... | 27 |
| | Context | 27 |
| | Participants | 28 |
| | School Selection | 28 |
| | Proximity Criteria..... | 29 |
| | Size, Grade Level, Scheduling, and Data Criteria..... | 29 |
| | Disclosure Criteria..... | 29 |
| | Summary of School Selection Criteria..... | 30 |
| | Other Criteria Considered | 31 |

| CHAPTER | Page |
|---|------|
| Criteria Considered but not Used | 31 |
| Recruitment of LEA Participants | 32 |
| Local Education Agency 1 | 32 |
| Local Education Agency 2 | 34 |
| Local Education Agency 3 | 35 |
| Summary of LEAs Participating in the Study | 36 |
| Individual Participants..... | 36 |
| Campus Administration Participants..... | 36 |
| Instructional Staff Participants | 37 |
| Student Participants..... | 37 |
| Summary of Individual Participants..... | 38 |
| Instrumentation..... | 38 |
| System to Enhance Educational Performance..... | 39 |
| Third Grade Screening Measure | 39 |
| Administration Instructions for STEEP3M..... | 40 |
| Administration of STEEP3M..... | 40 |
| Administration Time of STEEP3M..... | 41 |
| Features of STEEP3M Instrument | 41 |
| Section 1: Fill in Missing Numbers..... | 41 |
| Section 2: Complete the Pattern | 41 |
| Section 3: Write the Number in Standard Form..... | 42 |
| Section 4: Fractions of Pie Charts | 42 |
| Section 5: Fraction Comparisons | 42 |
| Section 6: Using Data..... | 42 |
| Section 7: Measurement | 43 |
| Section 8: Multiplication and Division | 43 |
| Teacher Survey Interview Questionnaire (T-SIQ) | 43 |
| Description of T-SIQ Instrument | 43 |
| Data Sources..... | 45 |
| Academic Excellence Indicator System (AEIS) | 45 |
| Classroom Grades (GRADE) | 46 |
| System to Enhance Educational Performance 3M (STEEP3M)..... | 46 |
| Texas Assessment of Knowledge and Skills (TAKS)..... | 46 |
| Teacher Survey Interview Questionnaire (T-SIQ) | 46 |
| Summary of Data Sources in the Study | 46 |
| Design | 47 |
| Data Analysis | 48 |
| Answering the Research Questions..... | 48 |
| Research Question One | 48 |
| Research Question Two | 49 |

| CHAPTER | | Page |
|---------|--|------|
| | Research Question Three | 50 |
| | Research Question Four | 51 |
| | Instrument Software for Data Analysis | 52 |
| | Organization of Data | 52 |
| | Software for Analysis | 52 |
| | Design Limitations | 53 |
| | Limitations of Sample | 53 |
| | Statewide Assessment Limitations | 53 |
| | Differentiation of the Study | 53 |
| IV | RESULTS | 54 |
| | Research Question One | 54 |
| | Research Question Two | 54 |
| | Research Question Three | 54 |
| | Research Question Four | 55 |
| | Participants and Setting | 55 |
| | Descriptive Information on Data Sources | 57 |
| | Descriptive Information on the STEEP3M | 58 |
| | Central Tendency of STEEP3M | 58 |
| | Variability of STEEP3M | 58 |
| | Distribution of STEEP3M | 58 |
| | Descriptive Information on the GRADE | 60 |
| | Central Tendency of GRADE | 60 |
| | Variability of GRADE | 60 |
| | Distribution of GRADE | 60 |
| | Descriptive Information on TAKS | 61 |
| | Central Tendency of TAKS | 61 |
| | Variability of TAKS | 62 |
| | Distribution of TAKS | 62 |
| | Summary of Information on Data Sources | 62 |
| | Research Question One | 63 |
| | Internal Consistency | 65 |
| | Item Difficulty | 65 |
| | Item Discrimination | 67 |
| | Cronbach's Alpha | 68 |
| | Criterion-related Validity | 68 |
| | Differential Predictability of TAKS by STEPP3M | 69 |
| | Differential Predictability of GRADE by STEPP3M | 71 |
| | Summary of Research Question One | 72 |

| CHAPTER | Page |
|---|------|
| Research Question Two | 72 |
| Teacher Perception of the STEEP3M for RtI..... | 78 |
| Summary for Research Question Two | 78 |
| Research Question Three | 79 |
| Content Analysis | 80 |
| STEEP3MContent Alignment to Curriculum | 80 |
| Classical Item Analysis | 82 |
| Correlation Analysis of Sub-groups | 84 |
| Correlation between STEEP3M and TAKS..... | 85 |
| Face Validity | 86 |
| Information from Instructional Staff | 87 |
| Teacher Reported Methods for Decision-making .. | 87 |
| Mathematics Textbook and Commercial Materials | 88 |
| Initial Perception of STEEP3M..... | 89 |
| Summary of Teacher Perception of STEEP3M | 89 |
| Summary of Research Question Three..... | 90 |
| Research Question Four | 91 |
| Administration of the STEEP3M | 92 |
| Time Limits of Administration | 92 |
| Information from Instructional Staff | 94 |
| Use of Released TAKS | 94 |
| Relationship of STEEP3M Score and Time to Complete .. | 94 |
| Scoring | 95 |
| Summary of Research Question Four | 95 |
| Summary of Results | 96 |
| V DISCUSSION AND SUMMARY | 97 |
| Purpose of the Study | 97 |
| Conduct of the Study | 98 |
| Research Results | 101 |
| Research Question One | 101 |
| Research Question Two | 102 |
| Research Question Three | 104 |
| Research Question Four | 106 |
| Research Summary..... | 107 |
| Research Limitations..... | 109 |
| Design Limitations | 109 |
| Timeline of Study..... | 110 |
| Statewide Assessment Limitations..... | 110 |
| Time Limit on Administration | 111 |
| Summary Conclusions..... | 111 |

| | Page |
|-------------------|------|
| NOMENCLATURE..... | 114 |
| REFERENCES | 116 |
| APPENDIX A | 132 |
| VITA | 133 |

LIST OF FIGURES

| FIGURE | | Page |
|--------|--|------|
| 4.1 | Histogram of Number of Items Correct for STEEP3M | 59 |
| 4.2 | Histogram of Math End of Year Grade (GRADE) | 61 |
| 4.3 | Histogram of TAKS Scale Score for Students in Study..... | 62 |
| 4.4 | Scatter Plot of TAKS versus STEEP3M Scores | 70 |
| 4.5 | Scatter Plot of GRADE versus STEEP3M Scores | 71 |
| 4.6 | Plots Section of STEEP3M Time Finished in Minutes..... | 93 |

LIST OF TABLES

| TABLE | Page |
|--|------|
| 3.1 Summary of School Selection Criteria..... | 30 |
| 3.2 Student Composition Percentages 2004-2007..... | 33 |
| 3.3 Third Grade Students Who Met TAKS Standard for Math 2004-2006 | 34 |
| 3.4 Student Enrollment Count Data Multi-Year History for 2005-2007 | 34 |
| 3.5 Summary of Participants Eligible for Study | 38 |
| 3.6 Indicators Used in Developing T-SIQ Questionnaire | 44 |
| 3.7 Summary of Measurements for the Study..... | 45 |
| 3.8 Timeline for Study | 47 |
| 3.9 Data analysis Plan for Research Question One (Psychometric Features).. | 49 |
| 3.10 Data analysis Plan for Research Question Two | 50 |
| 3.11 Data analysis Plan for Research Question Three | 51 |
| 3.12 Data analysis Plan for Research Question Four | 52 |
| 4.1 Summary of All Participants in Study..... | 56 |
| 4.2 Summary of Student Count by Data Source | 57 |
| 4.3 Descriptive Statistics Summary for STEEP3M, GRADE, and TAKS | 63 |
| 4.4 Item Analysis for STEEP3M: Average Item Difficulty | 66 |
| 4.5 Item Analysis for TAKS: Average Item Difficulty..... | 67 |
| 4.6 Item Analysis for STEEP3M: Average Item Discrimination..... | 68 |
| 4.7 Classification Accuracy Table for Prediction of whether Student Met TAKS Standard | 76 |
| 4.8 Selected Optimized Solution..... | 77 |
| 4.9 Summary of Number of STEEP3M Test Items to TEKS Objectives | 82 |
| 4.10 Descriptive Statistics on STEEP3M Sub-groupings by TEKS Objective.. | 83 |
| 4.11 Correlation between STEEP3M and TAKS by TEKS Objective | 86 |

CHAPTER I

INTRODUCTION

Mathematics is Fundamental

Increasing student performance on mathematics state standards is a national concern and a major goal of the American educational system (Clarke & Shinn, 2004; Jordan, Kaplan, Locuniak, & Ramineni, 2007; National Educational Goals Panel, 2002). Improving student performance is a priority, yet assessments indicate that mathematics is not learned well enough for children in the United States to compete internationally (National Research Council [NRC], 2006). Results of the 1996 National Assessment of Educational Performance (NAEP) indicated that only 21% of fourth-grade students performed at or above grade level proficiency in mathematics (Clarke & Shinn, 2004; Reese, Miller, Mazzeo, & Dossey, 1997). In 2003 fewer than 33% of fourth graders demonstrated proficient skills on the NAEP mathematics test (Manzo & Galley, 2003; VanDerHeyden & Burns, 2008). Current mathematics reforms emphasize utilizing rigorous mathematics standards for students with and without disabilities (National Council of Teachers of Mathematics [NCTM], 2007a). Considering the need to improve student performance and meet current standards, the importance of measuring and developing mathematical literacy for all students has never been greater.

In 2001, the Elementary and Secondary Education Act (ESEA, 1965) was reauthorized as the No Child Left Behind (NCLB) Act. As a result of new policy mandates under NCLB, expectations and demands on public schools in the United States have greatly increased (U.S. Department of Education [USDOE], 2002, 2003, 2005, 2006a). These mandates require improved student performance on assessments aligned to enrolled grade level (EGL) state standards for all children. Failure to demonstrate student proficiency on standards-based assessment, known as adequate yearly progress (AYP) under NCLB (2001), has potential consequences, including closure or restructuring for low-performing schools (USDOE, 2006a).

This dissertation follows the style of *Exceptional Children*.

Schools that have good results for students without special needs also have good results for students with special needs, including English language learners, students with disabilities, and those at-risk of failure (Malgrem, McLaughlin, & Nolet, 2005). As a result, there has been a move towards more inclusion in the educational classroom setting (Kunsch, Jitendra, & Sood, 2007; Malgrem, McLaughlin, & Norlet, 2005), requiring teachers to address the diverse needs of students at-risk of failure or with learning differences or disabilities (Carnine, Jones, & Dixon, 1994; Tomlinson et al., 1997). In this context of education, inclusion is a term that refers to the practice of educating students with special needs in the general education classes with the general education curriculum instead of in separate classes, for the whole day or a great portion of it.

In addition to NCLB, a special education policy mandate, the Individuals with Disabilities Education Improvement Act (IDEA) of 2004, contains explicit provisions for students with disabilities to have access to and make progress in the general state EGL standards curriculum. Both NCLB 2001 and IDEA 2004 call for larger numbers of students with disabilities to be included in EGL standards-based assessment (Lembke & Foegen, 2009; NCLB & IDEA Rule, 34 C.F.R § 200, 300, 2007).

Employment Outlook

According to the *Occupational Outlook Handbook*, basic mathematics, science, and technology skills are increasingly necessary for employment in today's world economic marketplace (Clarke & Shinn, 2004; Institute for Educational Leadership [IEL], 2003; U.S. Department of Labor [USDOL], 1997, 2007). In 1997, the U.S. Bureau of Labor Statistics reported that on average, 28-year-old workers who tested in the top quartile of mathematics skills on the NAEP earn 37% more than those in the lower quartiles (Clarke & Shinn, 2004; Riley, 1997; USDOL, 1997). The demand for students skilled in mathematics, along with the low-performance statistics, points to the need for studies on ways to increase student achievement in mathematics. There is a critical need for research in order to target early intervention and plans for formative evaluation, based on state standards curriculum components for increasing student

achievement (Clarke & Shinn, 2004; National Mathematics Advisory Panel [NMAP], 2008).

Problem Statement

The prevalence of mathematics learning disabilities in the general population, approximately 6% to 7% (Dirks, Spyer, vanLieschout, & Sonnevile, 2008; Murphy & Mazzocco, 2008), is comparable to that for reading disabilities. Approximately 5% to 8% of U.S. public school children exhibit characteristics of a learning disability (Badian, 1983; Fletcher, 2005; Fuchs et al., 2007; Geary, 2004; Gross-Tsur, Manor, & Shalev, 1996; Kunsch, Jitendra, & Sood, 2007; Lembke & Foegen, 2009; Lewis, Hitch, & Walker, 1994; Montague, 2007; National Center for Education Statistics [NCES], 2006). Researchers report that students with mild learning disabilities experience difficulty in most aspects of mathematics and progress at a slower rate than normally achieving peers (e.g., Carnine, Jitendra, & Silbert, 1997; Cawley & Miller, 1989; Englert, Culatta, & Horn, 1987; Harris, Miller, & Mercer, 1995; Mastropieri, Bakken, & Scruggs, 1991; Parmar, Cawley, & Frazita, 1996; Zentall, 1990; Zentall & Ferkis, 1993). The pattern of low performance in students with mild disabilities parallels the pattern of general education students in the U.S. who also experience difficulty with many mathematical concepts and problem solving (Carpenter, Matthews, Lindquist, & Silver, 1984; Kelley, Hosp, & Howell, 2008; NAEP, 1992).

Research on Mathematics

Even though the incidence of mathematics disabilities (MD) is similar to the incidence of reading disabilities (RD), less systematic research has been conducted on MD (Fuchs et al., 2007; Fuchs et al., 2008; Gross-Tsur, Manor, & Shalev, 1996; Lewis, Hitch, & Walker, 1994). Fuchs and Fuchs (2007) reviewed nine research studies on curriculum based measurement (CBM) that focused on the predictive utility of screening measures for forecasting MD at the end of second grade and the predictive validity of mathematics progress-monitoring tools. None of the studies reviewed reported predictive and correlation data aligned with the state standards-based assessment or decision utility data for lesson planning.

Previous studies on reading skills have identified similar critical needs for reading assessments designed to discover patterns of low performance. As a result, the U.S. Congress established the National Reading Panel (NRP) in 1997 to review a large body of research on reading, leading to the identification of effective reading screeners (Jordan, Kaplan, Locuniak, & Ramineni, 2007; National Institute of Child Health and Human Development [NICHD], 2000). In turn, the use of these assessments has fostered the development of evidence-based reading interventions (Bus & van IJzendoorn, 1999). Review of related literature indicates there has not been as much systematic mathematics research on to identify valid screening for potential mathematics difficulties and develop effective evidence-based interventions (e.g., Baker, Gertsen, & Lee, 2002; Fuchs et al., 2007; Gross-Tsur, Manor, & Shalev, 1996; Jordan, Kaplan, Locuniak, & Ramineni, 2007; Lewis, Hitch, & Walker, 1994). Additionally, the focus of studies to date has been narrow, addressing basic mathematics facts or simple computation (Fuchs, Fuchs, & Hollenbeck, 2007; Fuchs, Fuchs, & Zumeta, 2008).

In 2006, the U.S. Congress established the National Mathematics Advisory Panel (NMAP), modeled after the NRP, in order to examine and summarize the scientific evidence related to the teaching and learning of mathematics (USDOE, 2006b). The NMAP issued its final report in March 2008. It included a review of standards and accountability needed to ensure that students are learning mathematics skills, as well as a review of the mathematics literature. The Panel recommended research on formative assessment for students in the elementary grades in order to provide information for teachers to design or individualize instruction. Despite these recommendations, effective mathematics screeners have not been developed, researched, and/or consistently utilized. Few studies have been conducted on predictive ability to provide data to inform instruction or plan interventions. Without effective mathematics screeners aligned to state standards curriculum, students with mathematics difficulties are less likely to be identified and provided with appropriate instruction and intervention. In addition to the NMAP report, a recent meta-analysis of mathematics interventions indicated that the studies reviewed failed to produce significant results, primarily because of inadequate

assessment tools to identify student needs in order to plan appropriate interventions (Maccini, Mulcahy, & Wilson, 2007).

Mathematics Curriculum Based Measurement

Previous studies indicate curriculum based measurement (CBM) is a well-established tool for formative assessment, and could potentially be used for other purposes such as a prediction of state standards test scores. However, to date there are a limited number of validity studies between CBM and states standards tests (e.g., Connell, 2005; Skiba, Magnusson, Martson, & Erickson, 1986; Martson, 1989; Putnam, 1989). Carnine and Granzin (2001) reported that in the context of increased accountability to improve student performance, there is a need for general and special educators to have access to practical and reliable information regarding assessment educational tools. Tindal and Parker (1991) identified four criteria that can be used to guide the search for useful assessment procedures: 1) consistent administration and reliable scoring; 2) ability to discriminate among students at different skill levels; 3) criterion validity; and 4) sensitivity to growth. Research is necessary to extend the previous work on mathematics CBM in relation to these criteria and performance on state standards-based assessments. Further research on decision utility data for lesson planning is also needed.

Data Collection Methodology

Dillman (1978) reported that there is no one best way to collect data. Schools were selected for participation in the study using convenience sampling of campuses in Southeast Texas with third-grade classrooms. The methodology for selecting schools and determining their willingness to participate in the study is described in Chapter III. For this unfunded study, the most inexpensive methods were utilized to develop selection criteria for schools. Data for the study was collected through onsite administration of a mathematics screening instrument, interviews with educators, and reviews of educational records (e.g., performance on state standards-based assessment, teacher-given end-of-course grades). The researcher travelled to each campus to collect the data. A sample of local education agencies (LEA, n=3), campuses (n=3), educators (n=22), and third-grade

students (n=337) was utilized for this study. Sample size was determined based on review of literature and research guidelines as described in Chapter III.

Purpose of the Study

Identifying students at-risk of failure to master mathematics state standards as early as possible is necessary in order to target and implement interventions to improve student performance. It is important to determine formative curriculum based measurement that is predictive of performance on state standards in order to identify appropriate interventions (Fletcher, 2005; VanDerHeyden & Witt, 2005). This study was undertaken to extend knowledge on how educators could identify at-risk students through a simple, reliable process.

This research will examine the accuracy of a brief assessment vehicle as an indicator of performance on standards-based assessment, provide information regarding the validity for planning instruction, and evaluate the ability of such a CBM to aid in screening students for a multi-tiered decision-making process. This research differs from previous studies in that it focuses on mathematics rather than reading, conducts assessment in the whole group setting in the general education classroom, and utilizes an assessment designed based on national standards.

System to Enhance Educational Performance (STEEP)

STEEP is a company that develops educational tools and provides training for schools on how to use their tools correctly. Commercial assessment materials are available to conduct screening, to use as benchmark assessments, and to monitor progress. The company uses a standard protocol approach to quickly identify the type of intervention needed in reading or mathematics for students not achieving benchmarks (STEEP, 2007b). STEEP has provided a copyrighted grade 3 focal point mathematics screener for this study (which will be further reviewed in Chapter III) and has granted permission for its use. STEEP's senior scientist, Joe Witt, Ph.D., reports that the screener is based on the NCTM focal points.

Research Questions

The questions posed for this study are as follows:

Research Question One. How well does the mathematics curriculum based measurement (CBM) assessment tool, the STEEP3M, reflect end-of-year student performance on state standards-based curriculum, as measured by the end-of-year high-stakes TAKS test (internal consistency and criterion-related concurrent validity)?

Research Question Two. To what extent does the STEEP3M provide information to support a multi-tiered decision-making process in the third grade school setting, such as for grouping of students and placement?

Research Question Three. To what extent does the STEEP3M permit teachers to generate detailed mathematics lessons and unit planning for whole-class and individual needs?

Research Question Four. How efficiently can the one-time, group administered STEEP3M be administered to third grade students (as indicated by time to administer, score, report results, and provide feedback to teachers)?

Organization of Study

This study is organized in the following manner. This chapter presents the problem statement, describes the study's purpose, and lists research questions. Chapter II is a literature review that provides a rationale for development of the study's research questions. In Chapter III, the methodology for conducting the research is discussed. Research findings are reviewed in Chapter IV. Finally, Chapter V provides evidence for the research questions, utility of the data, limitations of the study, and recommendations for further study. Following Chapter V, the Nomenclature section provides information for names, acronyms, abbreviations, and terms used. The Appendix section contains supplementary materials.

CHAPTER II

REVIEW OF LITERATURE

Mathematic skills are essential to daily life, impacting a person's ability to function at home and work, and in the community. Although research on reading ability has received much focus in the U.S. in recent years, many students struggle with mathematics. Approximately 5% to 8% of school children have a mathematic disability (MD) (Jitendra, 2005). One major goal of the American education system is to increase student performance in mathematics (Clarke & Shinn, 2004; Jordan, Kaplan, Locuniak, & Ramineni, 2007; National Educational Goals Panel, 2002). In 1989, Anrig and LaPointe reported that only 16% of eighth grade students in the U.S. had mastered the content of a typical eighth-grade mathematics textbook. According to the 1996 results of the National Assessment of Educational Performance (NAEP), only 21% of fourth grade students performed at or above proficiency level in mathematics (Clarke & Shinn, 2004; Reese, Miller, Mazzeo, & Dossey, 1997). In 2003, fewer than 33% of fourth graders demonstrated proficiency on the NAEP mathematics test (Manzo & Galley, 2003; VanDerHeyden & Burns, 2008). This dissertation focuses on the utility of a formative mathematics curriculum based measurement (CBM) tool for early identification of at-risk students and to plan instruction to target student needs. This chapter will review literature on mathematics education, policy, and assessment tools.

Legacy of a Nation at Risk

In 1983, the publication of the report U. S. Department of Education (USDOE) publication of *A Nation at Risk* concluded that U. S. schools were mediocre at best and that the performance of American students was declining. Although these findings have been challenged, as a result of this report there has been much focus on improving student performance in U. S. schools (Ralph, 1994). In 2001, the Elementary and Secondary Education Act (ESEA) was reauthorized as the No Child Left Behind Act of 2001 (NCLB). The NCLB (2001) represents a change in and clarification of the relationship between the federal government and state/local education agencies

(SEA/LEA) regarding control of education in schools and classrooms (Simpson, LaCava, & Graner, 2004; Sunderman, Orfield, & Kim, 2006). These policy mandates call for improving student performance on state standards, which in turn have greatly increased expectations and demands on public schools in the U.S. (Jitendra, Sczesniak, & Deatline-Buchman, 2005).

Literature and research indicate that remediation efforts for struggling learners have not resulted in significantly improved outcomes, thus leading to a shift in focus to prevention models (Chard et al., 2005). The lack of evidence to support continuing remediation in separate learning environments has resulted in a movement toward inclusion of children with special needs in the general education classroom setting (Kunsch, Jitendra, & Sood, 2007; Malgrem, McLaughlin, & Nolet, 2005). With inclusion, teachers are increasingly required to address the diverse needs of students who are at-risk of failure, have learning differences or disabilities, or are English language learners (Carnine, Jones, & Dixon, 1994; Tomlinson et al., 1997). Findings from several studies support the inclusion movement (e.g., Bottge, Rueda, LaRoque, Serlin, & Kwon, 2007) and suggest students with MD can participate in learning activities. These studies also suggest that it is not necessary to wait until all related procedural skills are mastered to teach concepts for understanding.

In addition to the NCLB of 2001, the latest reauthorization of the special education law, the Individuals with Disabilities Education Improvement Act (IDEA) of 2004, includes explicit provisions for students with disabilities to have access to and make progress in the general state standards-based curriculum (Miller & Hudson, 2007). In an unprecedented alignment of general and special education legal mandates, policy-makers embedded IDEA with NCLB provisions, such as providing highly qualified teachers, research research-based intervention, and standards-based assessment.

Standards-based Assessment

The latest regulations of the NCLB of 2001 and IDEA 2004 include provisions that students be assessed utilizing enrolled grade level (EGL) standards-based assessment (NCLB & IDEA Rule, 34 C.F.R § 200, 300, 2007). Schools face significant

consequences, ranging from loss of federal funding, provision of supplemental services, school takeover, and restructuring (IDEA, 2004; NCLB, 2001; Shippen, Houchins, Calhoon, Furlow, & Sartor, 2006), if they fail to demonstrate student proficiency on standards-based assessment, known as adequate yearly progress (AYP). Under NCLB (2001), SEAs define AYP goals for LEAs, including a performance assessment of all students on grade-level standards. Data from student performance on the state standards-based assessment are disaggregated to reflect the impact of each grade level and subgroup (e.g., ethnicity, special education, English language learners) on AYP (Shippen, Houchins, Calhoon, Furlow, & Sartor, 2006). The performance of a small number of students may impact the success of an entire school campus or LEA.

Assessment for Early Identification. Most educators agree that it is important for schools to assess children's early mathematics abilities in order to predict later performance on school tasks (Savage & Carless, 2004). Under NCLB 2001, summative assessment of students on state standards and determinations of AYP begins in third grade. The NCLB also requires assessment in reading and math of all students in grades three through eight. States and schools across the U. S. generally conduct formative early screening of students to identify potential areas of mathematics difficulties in order to target interventions and plan instruction (Gertsen & Jordan, 2005; Jordan, Kaplan, Locuniak, & Ramineni, 2007). Early screening generally begins in kindergarten, first, or second grade, after children have received instruction on mathematic academic skills. In order to plan and deliver instruction in the general education classroom setting, educators regularly examine the way student needs are identified (Kunsch, Jitendra, & Sood, 2007).

Research on CBM. CBM is a well-established tool for formative assessment (Busch & Espin, 2003) and could potentially be used for other purposes, such as a prediction of state standards test scores. Although this potential for CBM exists, there are only a limited number of validity studies between CBM and state tests (e.g., Connell, 2005; Fuchs, Fuchs, & Zumeta, 2008; Komatsu, 2004; Skiba, Magnusson, Martson, & Erickson, 1986; Martson, 1989; Putnam, 1989). Carnine and Granzin (2001)

reported that in the context of increased accountability to improve student performance, there is a need for general and special educators to have access to practical and trustworthy information regarding educational assessment tools. Although there are other evaluation methods for gaining this information, researchers have proposed that CBMs are more cost-effective, quicker and easier to administer, more sensitive to student growth, and allow for continual progress monitoring (Ardoin, Sulto, Witt, Aldrich, & McDonald, 2005; Lembke, Deno, & Hall, 2003; VanDerHeyden, Witt, Naquin, & Noell, 2001; Ysseldyke & Algozzine, 2006).

Economic Factors. A driving force behind recent educational reforms has been the demand for skilled workers who can apply problem-solving skills (Fuchs, Fuchs, & Courey, 2005). Mathematics, science, and technology skills are increasingly necessary for employment in today's world, according to the current *Occupational Outlook Handbook* (Clarke & Shinn, 2004; IEL, 2003; USDOL1997; 2007). Individuals with mathematics fluency have greater earning potential. The Bureau of Labor Statistics reported that on average, 28-year-old workers who tested in the top quartile of mathematic skills on the NAEP earn 37% more than those in the lower quartiles (Clarke & Shinn, 2004; Riley, 1997; USDOL, 1997). Demonstrating competency in mathematics involves both the acquisition of basic mathematic skills and the ability to integrate those skills into problem solving (Fuchs & Fuchs, 1996; Fuchs, Fuchs, & Courey, 2005).

Policy Reforms. The NCLB Act of 2001 mandates accountability for standards-based performance through school reform requirements, which are designed to support a student's AYP. Under the latest NCLB 2001 and IDEA 2004 regulations, 97% to 99% of all students must be assessed using state standards-based curriculum for their EGL. The remaining students may be assessed on alternate academic standards or modified academic standards, which are aligned with the state's academic content for their EGL. Therefore, all students are currently evaluated by EGL standards-based assessment. To meet these policy mandates, the U.S. DOE (2002) has identified school reform model characteristics to identify student-specific patterns of strengths and weaknesses. Schools

are required to implement programs that plan for evaluation of strategies and integrate comprehensive designs that have aligned components.

Students with Disabilities under IDEA 2004. The first special education law, known as the Education for All Handicapped Children Act, was passed in 1975 and reauthorized as IDEA in 1997 and 2004. Since the original 1975 legislation, there has been a consistent increase in the number and proportion of children with learning disabilities being served in special education programs. Between the 1976-77 and 2003-04 school years, the incidence of public school students with disabilities in all categories grew from 8.3% to 13.7%. During this same period, the incidence of students identified as having specific learning disabilities grew from 1.8% to 5.8%. (NCES, 2006). IDEA 2004 provides for eight subcategories of specific learning disabilities, including two areas of mathematics (calculation and mathematics problem solving) and three areas for reading (basic reading, reading comprehension, and reading fluency) (IDEA 2004). National statistics are neither available nor disaggregated within the eight subcategories of learning disabilities (Calhoon, 2008). In general, only primary disabilities are reported, which further confounds more specific estimates of the number of students identified as having MD under IDEA 2004. The incidence of learning disabilities in mathematics is reported to be similar to the incidence of learning disabilities in reading disabilities (RD) (Dirks, Spyer, vanLieschout, & Sonnevile, 2008; Gross-Tsur, Manor, & Shalev, 1996; Fletcher, 2005; Fuchs et al., 2007; Gross-Tsur, Manor, & Shalev, 1996; Lewis, Hitch, & Walker, 1994; Kunsch, Jitendra, & Sood, 2007; Lembke & Foegen, 2009; Lewis, Hitch, & Walker, 1994; Montague, 2007; Murphy & Mazzocco, 2008), yet less attention has been devoted to understanding MD (Geary, 2004; Manzano & Galley, 2003; Murphy & Mazzocco, 2008).

Most students with mild learning disabilities experience difficulty in all aspects of mathematics and progress at a slower rate than normally achieving peers (e.g., Carnine, Jitendra, & Silbert, 1997; Cawley & Miller, 1989; Englert, Culatta, & Horn, 1987; Harris, Miller, & Mercer, 1995; Mastropieri, Bakken, & Scruggs, 1991; Parmar, Cawley, & Frazita, 1996; Zentall, 1990; Zentall & Ferkis, 1993). This pattern of low

performance for students with mild disabilities parallels the pattern of general education students in the U.S. who are not identified for special education or related services, but who may also experience difficulty with mathematical concepts and problem solving (Carpenter, Matthews, Lindquist, & Silver, 1984; NAEP, 1992). Review of related literature suggests that these differences can be attributed to the amount and kind of exposure to mathematics instruction rather than real differences in student abilities (Stevenson, Chuansheng, & Lee, 1993; Stevenson et al., 1990; Stigler, Lee, & Stevenson, 1990). Unlike reading, poor achievement in mathematics may actually worsen as a student progresses through school, primarily due to the cumulative nature of mathematics (Montague, 2007). In other words, new mathematic skills depend on mastery of previous concepts and skills. In addition to students with identified mathematics disabilities, many students fail to develop mathematical skills, with disability levels ranging from performance below expected levels, to dyscalculia.

National Reading Panel. The U.S. Congress established the NRP in 1997 to review research on reading (Jordan, Kaplan, Locuniak, & Ramineni, 2007; NICHD, 2000; Snow, Burns, & Griffin, 1998). The NRP reviewed information from approximately 100,000 research studies published since 1966, oral and written testimony from 125 individuals, and five public hearings. As a result of the Panel's work, two federally sponsored reviews of literature (Snow, Burns, & Griffin, 1998; NRP, 2000) were created to document the effectiveness of systematic, early literacy interventions (Otaiba & Fuchs, 2006). The findings of these two reports were referenced in NCLB (2001) and IDEA (2004) with specific mention of *scientifically based reading research* such as the Reading First initiative. Further, review of research suggested that reading problems might be preventable for the majority of students if they received early intervention (Allington, 1994; Ardoin et al., 2004; Ardoin, Sulto, Witt, Aldrich, & McDonald, 2005; Goldenberg, 1994; Hiebert & Taylor, 1994; Pikulski, 2002; Reynolds, 1991). Researchers and educators note that as many as 30% of children are at-risk of experiencing reading difficulties (Otaiba & Fuchs, 2006). Since the incidence of MD is

similar to RD, these numbers may parallel children's risk for experiencing mathematics difficulties.

National Mathematics Advisory Panel. The NMAP, modeled after the NRP, was established in 2006 in order to examine and summarize the scientific evidence related to the teaching and learning of mathematics. Following the path of the NRP, the NMAP sought information on what was known about early identification, interventions, and remediation for mathematics. Issues studied by the NMAP included a review of standards and accountability needed to ensure that students are learning mathematic skills. Over two years, the panel reviewed 16,000 research studies and related documents, public testimony from 110 individuals, written commentary from 160 organizations and individuals, input from 12 public meetings, and survey results from 743 Algebra I teachers. The Panel produced a final report in March 2008 that synthesized existing research and offered 45 recommendations on mathematics education (NMAP, 2008).

Overall the NMAP (2008) report recommended streamlining the curriculum in kindergarten through grade eight, improving the quality of NAEP and state assessments, and emphasizing the development of student skills and knowledge leading to algebra. The report recommended that teachers utilize formative assessment to design and individualize instruction. The NMAP also noted that although research to date on formative assessments for mathematics was limited, it was sufficiently promising to recommend continuing this inquiry to identify assessments developed from sampling major curriculum objectives and state standards. Additionally, according to the NMAP, both small-scale experiments on the basic science of learning, as well as large-scale randomized experiments examining effective classroom practices, are needed to address important questions in mathematics education. In order to target and implement appropriate interventions for struggling students, it is important to determine formative curriculum based measurements that are predictive of performance on state standards (Fletcher, 2005). Most of the reviewed formative mathematics assessments demonstrate criterion-related validities in the 0.5 to 0.7 range (Foegen, Jiban, & Deno, 2007).

According to the NMAP (2008), although these validities are weaker than those found in reading, they appear to be reasonable.

A separate meta-analysis of studies on mathematics interventions reported that many studies reviewed failed to produce significant results, due to assessment tools that are inadequate for identifying student needs in order to plan appropriate interventions (Maccini, Mulcahy, & Wilson, 2007). Without effective mathematics assessment screeners, students with difficulties are less likely to be identified and be provided appropriate instruction and intervention.

Initial Feedback on the NMAP Report. The NMAP report, although well received, has been criticized for adopting a strict and narrow definition of scientific evidence, employing methodological bias in selection of research studies, and constructing definitions of teaching that are not reflective of practices in U.S. classrooms. The report summarized each subpanel's report, but to date has not integrated the nearly 900 pages of task group and subcommittee findings (Boaler, 2009; Cobb & Jackson, 2009; Confrey, Maloney, & Nguyen, 2009; Greeno & Collins, 2008; Kelly, 2009; Lobato, 2009; Shepard, 2009; Thompson, 2009). The final report is primarily a listing of 45 findings and recommendations.

Additional Information on Mathematics Research. As noted, there has been less systematic research conducted on MD than on RD (Fuchs et al., 2007; Gross-Tsur, Manor, & Shalev, 1996; Lewis, Hitch, & Walker, 1994). Calhoon (2008) reported that a review of literature yielded 578 articles on the development, uses, and implementation of CBM. Of these, only 32 were in the area of mathematics. Fuchs and Fuchs (2007) reviewed nine research studies that were conducted on kindergarten or first grade, reporting predictive validity including mathematics outcomes. They observed that most of the studies used time-consuming screeners that measured a narrow range of skills. Additionally, none of these studies reported on both correlation data predictive of performance on the state standards-based assessment as well as decision utility data for lesson planning.

Defining Mathematics. The term mathematics is derived from the Greek term for learning, study, and science, and has evolved to refer to the body of knowledge about quantitative concepts, including quantity, structure, space, and change. Knowledge and use of mathematical tools and symbols have always been integral to both individual and community life. From prehistoric times, humans have used mathematical representations for abstract concepts such as time, days, or seasons. The field of mathematics has evolved from simple tally recording of knotted strings for counting to the study of relationships using numbers, symbols, shapes, and quantities (Oxford, 2007). This evolution has been viewed as developing from the necessity to represent applied or abstract concepts, including economic issues like taxation, commerce, and land measurement, and the prediction of astronomical events. Mathematics conveys concepts using symbol association, just as reading does. However, symbol association for mathematics is often complex and difficult for students to master. Mathematics requires a structured sequence of concepts that must be in place to build upon, similar to establishing a building's foundation before the walls are raised. The learning process is further complicated by the nature of mathematics, with computations generally resulting in an answer that is either correct or incorrect. Mathematics includes the disciplines of arithmetic, algebra, calculus, geometry, and trigonometry (NCTM, 2007a).

Multi-tiered Support Systems for Students. The latest legislative mandates require implementing a multi-tiered system of support designed to provide all students with standards-based curriculum, high-quality instruction, and scientifically validated or research-based interventions. These multi-tiered systems of support have recently become known as response to instruction or intervention (RtI) systems. RtI systems include screening of all students, targeting individual needs, providing additional research-based interventions, and monitoring progress (Fuchs et al., 2007; Colorado DOE, 2007; Fuchs et al., 2008; New Mexico DOE, 2006; Washington DOE, 2007; West Virginia DOE, 2007).

Criteria for Useful Assessment Procedures. In 1991, Tindal and Parker identified four criteria that can be used to guide the search for useful assessment procedures: 1) consistent administration and reliable scoring; 2) ability to discriminate among students at different skill levels; 3) criterion validity; and 4) sensitivity to growth (Lembke, Deno, & Hall, 2003). Research to extend the previous work on mathematics CBM in relation to these criteria and performance on state standards-based assessment is necessary. Further research on decision utility data for lesson planning is also needed.

Curriculum Based Assessments. Curriculum based assessment (CBA) is defined as a set of procedures using direct observation and recording of student performance in the local standards-based curriculum as the basis for gathering information in order to plan instruction (Deno, 1985, 2003; King-Sears, 1994). Research has been conducted on the use of CBA to measure student performance in the basic academic skill areas of reading, writing, spelling, and mathematics; however, mathematics has been the least represented in literature (Thurber, Shinn, & Smolkowski, 2002). Diagnostic measures to identify specific difficulties or disabilities in mathematics do not exist (Chard et al., 2005; Geary, 2004).

Models of CBA. After reviewing literature and research, Connell (2005) reported that the dearth of CBA applications for mathematics might be a function of the differences among the CBA models that have emerged. Connell (2005) identified three main CBA models: 1) accuracy model of CBA (Gickling & Havertape, 1981; Gickling & Thompson, 1985; Hargis, 1987); 2) criterion-referenced assessment (CRA) model (Blankenship, 1985; Idol-Maestas, 1983); and 3) fluency model (Deno & Mirkin, 1977). For the purposes of this study, the fluency model, also known as curriculum based measurement (CBM), will be used. CBM differs from CBA and CRA in that the purposes of CBM include use as a direct assessment of student performance for student progress monitoring (Christ & Vining, 2006; Good & Kaminski, 1996; Idol, Nevin, & Paolucci-Whitcomb, 1999; Ysseldyke & Algozzine, 2006); to design instruction (Shapiro, 1996); and to document gains in academic skills (Martson, Deno, & Kim, 1995).

Differentiated Features of CBM. CBM is considered to be an “authentic” way to assess student performance and has been used as an evaluative tool in the educational setting for over 20 years (Shapiro, Edwards, & Zigmond, 2005; Shinn, 1989). CBM is an assessment vehicle that is characterized by a set of standard directions, a timing device, a set of materials, scoring rules, standards for judging performance, and record forms (Christ, Scullin, Tolbize, & Jiban, 2008). The student is asked to do an activity that is similar to familiar classroom activities. In fact, CBM often looks like a classroom teaching activity, but provides evaluation information. One differentiating feature is that CBM can align with classroom activities or standards-based curriculum, and therefore can more accurately assess what is taught (Hosp, Hosp, & Howell, 2007). The CBM content must be a subset of the content domain of a specific curriculum (Helwig, Anderson, & Tindal, 2002). Therefore, performance on a CBM is representative of the student’s mastery level of curriculum content (Calhoon, 2008). Good and Jefferson (1998) conducted correlational studies between CBM measures and various measures of achievement for reading, writing, and mathematics, finding the correlation was over 0.60. In a study by VanDerHeyden, Witt, and Naquin (2003), a CBM screener highly correlated (0.545) to the Iowa Test of Basic Skills (ITBS) scores in reading and mathematics.

Additional features differentiate CBM from other forms of assessment. The CBM is standardized, in that the behaviors to be measured are specified, as are the procedures for measuring those behaviors; its focus is long-term, so the testing methods and content remain consistent; and it reflects the performance desired at the end of the year (Fuchs, Fuchs, Hosp, & Hamlett, 2003; Hosp & Hosp, 2003; Hosp, Hosp, & Howell, 2007).

Formative Assessment. Educators generally agree that it is important for schools to identify areas of student abilities that might predict later performance on curricula and school tasks (Helwig, Anderson, & Tindal, 2002; Methe, Hintze, & Floyd, 2008; Savage & Carless, 2004). This is known as formative assessment. It can be used to identify the existence and type of student needs as early as possible in order to plan instruction and intervention. Researchers and educators can administer formative assessments to class or

grade-level groups and use them as a diagnostic baseline to target skills for planning classroom lessons or individual instruction (Fuchs & Fuchs, 1986; Fuchs, Fuchs, & Hamlett, 1989; Fuchs, Fuchs, Hamlett, & Stecker, 1990; Ketterlin-Geller, McCoy, Twyman, & Tindal, 2003; NMAP, 2008).

CBM as Formative Assessment. CBM has been documented as a well-established tool for formative assessment (Connell, 2005). Once students have been taught academic skills, targeted assessment to determine skill acquisition and mastery may be administered. Academic skills are introduced in kindergarten through second grade; therefore, third grade is generally targeted for the first administration of statewide assessments. Kunsch, Jitendra, and Sood (2007) report that research developed CBMs are typically more sensitive to student performance than standardized assessments, due to their match to curriculum content. Identifying CBMs that can be administered quickly to whole groups and that are tied to state or national standards may provide useful information to teachers for planning instruction (Busch & Espin, 2003; Ketterlin-Geller, McCoy, Twyman, & Tindal, 2003; VanDerHeyden, Witt, Naquin, & Noell, 2001). The National Center on Student Progress Monitoring (NCSPM, 2007) reported that there is emerging evidence that CBM may be used to predict performance on state assessments.

Summative Assessment. Summative assessment occurs after instruction and a period of work in order to summarize the development or proficiency of the learner. In addition to providing information on a student's performance, summative tests also can be used diagnostically to identify patterns of strengths and weaknesses. State standards-based assessment is generally summative, occurring at the end of the school year, typically too late for planning current year instruction. In simple terms, formative assessment is used to plan for learning and summative assessment is used to gauge the extent of learning (Butler & McMunn, 2006).

Standards-based Curriculum. Carnine, Jitendra, and Silbert (1997) suggest that curriculum should be guided by several basic principles. In particular, curriculum should be organized around fundamental ideas, concepts, or principles to increase understanding of complex content. The fundamental ideas provide a framework for

organizing instructional sequencing to help students make connections within and among subject areas such as mathematics. Once fundamental ideas have been identified, a series of specific activities and lessons can be designed for instruction.

National Council of Teachers of Mathematics (NCTM). The NCTM (1989) developed standards for teaching mathematics. Instructional programs that make explicit connections between related skills and concepts are more likely to facilitate learning (Carnine, Jitendra, & Silbert, 1997) and appropriate generalizations (Carnine, Jones, & Dixon, 1994; NCTM, 1989). In 1991 the NCTM published *Professional Standards for Teaching Mathematics*, which described the elements of effective mathematics teaching. In 1995, the NCTM published the *Assessment Standards for School Mathematics*, listing objectives against which assessment practices can be measured. Analyzing what to teach in the classroom can help determine when and how to teach specific content. Student assessment tools that are criterion-referenced using the NCTM standards have been found to be particularly relevant in the case of students with mild disabilities.

As a part of the mathematics and standards-based reform movements, the NCTM published the *Principles and Standards for School Mathematics* in 2000 (Xin, Jitendra, & Deatline-Buchman, 2005). It revised the three previous documents and created a single resource to be used for improvement of mathematics curricula, teaching, and assessment. This document includes content and process standards for the knowledge and the skills students should learn, such as numbers and operations, algebra, geometry, measurement, data analysis, and probability. Currently, 46 of the 50 separate state-developed standards have been shaped or influenced by the NCTM standards (Berkas & Pattison, 2007; Burrill, 1997; Helwig, Anderson, & Tindal, 2002).

NCTM Focal Points. The NCTM has identified and described three curriculum focal points for each grade level, pre-kindergarten through eighth grade (NCTMa, 2007). These focal points, along with integration at each grade level and connections across grade levels, form a national comprehensive mathematics curriculum. In order to increase and strengthen students' use of mathematical processes, instruction in content areas should incorporate 1) the use of mathematics to solve problems; 2) an application

of logical reasoning to justify procedures and solutions; and 3) an involvement in the design and analysis of multiple representations to learn, make connections among, and communicate about the ideas within and outside of mathematics (Miller & Hudson, 2007). According to the NCTM (2007a), the purpose of identifying grade-level curriculum focal points is to enable students to learn the content in the context of a focused and cohesive curriculum that implements problem solving, reasoning, and critical thinking. These curriculum focal points identify major instructional goals and desirable learning expectations, and are not just a list of objectives for students to master (NCTM, 2007a). The NCTM (2007a) *Principles* call for instruction designed to increase a student's capacity to think and reason mathematically (Helwig, Anderson, & Tindal, 2002). Research is needed on the CBMs that were developed utilizing the NCTM's standards to determine whether they reflect performance on state standards-based assessments.

Curriculum Focal Points for Grade 3. For third grade, the NCTM (2007a) recommends the following three topical areas for content emphasis: numbers and operations and algebra; numbers and operations; and geometry. The NCTM recommends these focal points be addressed in the contexts that promote problem solving, reasoning, communicating, making connections, and designing and analyzing representations. Numbers, operations, and algebra encompass the development of the student's understanding and application of strategies for multiplication and division. Numbers and operations include the understanding of fractions and fraction equivalence. Geometry refers to the ability to describe and analyze properties of two-dimensional shapes. Since assessment of student progress on state standards generally begins at third grade, specific research is needed on CBM that were developed utilizing the NCTM third grade focal points, to determine whether the CBM is reflective of performance on state standards-based assessment and to identify specific skill needs to plan instruction.

Texas Essential Knowledge and Skills. In Texas, the Texas Essential Knowledge and Skills (TEKS) are the state-mandated standard curriculum guideline. The TEKS establish what every student should know and be able to do (Texas Administrative Code

[TAC], 2009). Similar to the NCTM standards, the TEKS have focal points for number, operation, and quantitative reasoning; patterns, relationships, and algebraic thinking; geometry and spatial reasoning; measurement; and probability and statistics. The TEKS are used by LEAs and teachers to guide curriculum decisions, select instructional materials, and plan lessons. Although separate from the NCTM standards, the TEKS are aligned with the NCTM (TEA, 2009c).

Screening of Student Performance. In a multi-tiered system of support for students, all students are initially screened with a universal screener. Universal screening has three basic levels of use. First, the screener provides information on individual learners, their skill abilities, and areas of need. This information helps student support teams and instructional staff target interventions to support students in specific areas of skill needs. Second, the screener provides information on whole-class abilities and performance. The screener needs to be aligned to state or national standards-based curriculum in order to provide sufficient information for lesson planning. Third, as an extension of the first two purposes, the screener needs to be reflective of student performance on standards-based assessment. Additionally, the screener must be designed to gain the maximum amount of information about individual and group performance, with a minimal intrusion on classroom instructional time. Studies are needed in the area of mathematics to identify valid screening instruments for potential mathematic difficulties and to develop effective evidence-based interventions (e.g., Ardoyn et al., 2004; Bryant, 2007; Bryant & Bryant, 2006a, 2006b; Fuchs et al., 2007; Gross-Tsur, Manor, & Shalev, 1996; Jordan, Kaplan, Locuniak, & Ramineni, 2007; Lembke, Foegen, Whittaker, & Hampton, 2008; Lewis, Hitch, & Walker, 1994).

Criterion-referenced Tests. Criterion-referenced tests (CRT) are designed with questions written according to a specific content within a subject area domain. CRTs are intended to measure how well a student has learned a predetermined set of skills within the content domain. They report how well students are doing, relative to a predetermined performance level on a specified set of educational goals or outcomes included in the state standards. Key skills identified for the mathematic domain in the third grade CRT

include number and operations, measurement, geometry, algebra, and data analysis (NCTM, 2007b). Generally, problems with CRT are that they are time-consuming and lengthy, and each test must be designed to sample the criterion fairly.

Standards-referenced Tests. Many states, including Texas, have developed assessments using state standards or curriculum frameworks to design standards-referenced testing (SRT) or standards-based assessment (Center for Public Education [CPE], 2007; Education Commission of the States [ECS], 2002; O'Neill & Stansbury, 2000). The SRTs link assessment to curriculum, are designed to compare the student's performance to a standard of achievement, and incorporate new forms of assessment, such as writing essays or solving real-life mathematic problems (ECS, 2002). Since SRTs are designed to test what the student has learned, they are considered to be a summative test. The emphasis for formative tests is to assess in order to determine how best to help the student. In addition to the state-adopted standards, states have set a performance standard for the SRTs that determine proficiency levels for subject areas. The SRT is then based on the state standard and the results are reported in terms of the proficiency levels. In Texas, the Texas Assessment of Knowledge and Skills (TAKS) is a form of CRT and also an SRT because it is designed to determine a student's mastery of a specified content as well as mastery of state standards-based curriculum (TEA, 2009c).

Statewide Assessment in Texas. Statewide assessment of students in Texas began in 1979 with the Texas Assessment of Basic Skills (TABS) test. At that time, Texas did not have a statewide curriculum; therefore, committees of Texas educators developed the test. In 1984, the Texas Legislature re-worded the Texas Education Code (TEC), requiring the assessment program to be changed to measure minimum skills rather than competence in basic skills. The test was redesigned and re-named the Texas Educational Assessment of Minimum Skills (TEAMS). In 1990, the Texas Legislature again changed state law to include the requirement for a new criterion-referenced program, and a new test was developed, the Texas Assessment of Academic Skills (TAAS). The TAAS test shifted the focus of the statewide assessment from minimum skills to academic skills,

which aligned with state curriculum known as the Essential Elements. The TAAS test was designed with the intent to assess high-order thinking skills and problem solving in mathematics, reading, and writing. In 1993, the Legislature enacted the creation of a statewide accountability system that included public reporting of ratings based on TAAS performance for campuses and LEAs. In order to facilitate comparison of performance across grade levels, the Texas Learning Index (TLI) was developed in 1994. The TLI was designed to help LEAs determine whether individual students were making the AYP necessary in order to meet minimum requirements on the exit-level test (TEA, 2009c).

Texas Assessment of Knowledge and Skills. In 1999, the statewide test again was redesigned to increase rigor and renamed the Texas Assessment of Knowledge and Skills (TAKS). TAKS was aligned to the state curriculum, the TEKS. The TAKS was field-tested in 2002 (grades 3-11) and first administered in the spring of 2003. The TAKS tests reading in grade 3 and reading and mathematics in grades 5 and 8. Students are required to demonstrate proficiency on the TAKS and achieve passing grades in order to advance to the next grade level. In eleventh grade students must pass the TAKS test (in reading, writing, mathematics, science, and social studies) and complete the required number of course credits in order to receive a high school diploma. The Texas Education Code (TEC) charges the Texas State Board of Education with establishing the passing standards (performance standards) on the TAKS test (TEA, 2009c). Performance on the TAKS is a major component of the state's accountability system under NCLB 2001, and is used to measure a student's AYP. Studies on the correlation between formative screening assessments and performance on TAKS are needed to help teachers identify skill needs, plan lessons, and target individual or whole-class interventions.

The TAKS test is designed to be a CRT in that it measures a student's proficiency within content domains; it is an SRT in that it compares the student's performance to a standard of achievement. The TAKS test is designed to be a summative test of a student's mastery of the TEKS, and therefore is a CBM. It is problematic to use TAKS for monitoring student progress because its administration is time-consuming and its questions numerous. Additionally, in progress monitoring, each test must fairly

sample the content domain. Since the TAKS is designed to measure a student's performance in a wide content domain, this precludes its ability to fairly sample the student's depth and breadth in the domain.

CBM and State Standards. CBM has been successfully used for formative assessment with students. Once students have begun explicit instruction in mathematics, generally in the first or second grades, it is possible to identify students who are at-risk (Clarke & Shinn, 2004; Fuchs, Fuchs, & Zumeta, 2008). Since CBMs can be developed using state standards, the results have the potential to provide closer links for planning instruction, targeting intervention, and predicting performance on the state standards-based assessment. Recent research indicates that simple assessments matched the results of extended analysis or multiple probes approximately 83% of the time. This suggests the utility of conducting abbreviated assessments to differentiate skill deficits from performance deficits (Duhon et al., 2004). A few previous studies have been conducted to examine the accuracy of brief assessment in predicting student response to intervention. Further research is needed to determine how educators could simplify the processes of identifying at-risk students and planning for instruction by using results from brief, simple, group administered mathematics CBMs.

System to Enhance Educational Performance (STEEP). STEEP is a company that develops educational tools and provides training for schools on how to use these tools correctly. Commercial assessment materials are available to conduct screening and benchmark assessments, and to monitor progress. Products include kits comprised of materials for assessing skills such as oral reading fluency, maze (reading comprehension), and mathematics. The company uses a standard protocol approach to quickly identify the type of reading or mathematics intervention needed for students who do not achieve benchmarks (STEEP, 2007b).

Identifying Mathematics Problems

The inability to master mathematics calculations and problem solving contributes significantly to the rising incidence of student failure, referrals for special education evaluations, and dropout rates. In 1985, Carpenter noted that a student in a special

education classroom spent less than one-third of the time studying mathematics. Carnine, Jitendra, and Silbert (1997) reported that students with mild disabilities experience difficulty in most areas of mathematics. Further, this pattern of underperformance parallels poor student performance in the general education setting (Carpenter, Matthews, Lindquist, & Silver, 1985; Kelly, Hosp, & Howell, 2008), leading to the conclusion that the question is not whether to change mathematics education, but where to begin (Carnine, Jitendra, & Silbert, 1997).

Integral to determining a starting point for change is the identification of an accurate, brief curriculum-based measurement for mathematics in order to indicate performance on standards-based assessment, provide information regarding the validity for planning instruction, and evaluate the ability of such a CBM to aid in screening students for a multi-tiered decision-making process (Colorado DOE, 2007; NMAP, 2008; Shinn, 1989). Educators would benefit from the increased accuracy of predictive achievement scores and the discovery of measures that are quick and efficient (Helwig, Anderson, & Tindal, 2002; NMAP, 2008; VanDerHeyden, Witt, Naquin, & Noell, 2001). Research that differs from previous studies is needed: research that focuses on mathematics rather than reading, conducts assessment in the whole-group setting of the general education classroom, and utilizes an assessment design based on national standards (NMAP, 2008; VanDerHeyden & Witt, 2005).

CHAPTER III

METHOD

Context

This research was designed to examine the usefulness of a brief assessment that is aligned to national curriculum standards in order to predict student performance, identify students at-risk of failure, and plan instruction. Specifically, the aim of the study was to gather evidence on the usefulness of the *System to Enhance Educational Performance Grade 3 Focal Mathematics Assessment Instrument* (STEEP3M), a mathematics curriculum based measurement (CBM), as a formative, universal screener. This study focused on the STEEP3M assessment tool designed to identify students with or without disabilities who might respond to intervention (STEEP, 2007b). The timeline of the study targeted gathering data on the administration of the STEEP3M within close proximal time of the Spring, 2007 administration of the third grade statewide mathematics assessment, the Texas Assessment of Knowledge and Skills (TAKS). The timeline and design will be described in this chapter.

Since the study was on the use of universal screeners to predict performance on standards-based assessment, all third grade students who attended selected elementary campuses and took the regular state standards-based assessment for third grade mathematics, the TAKS, were eligible to participate in the study. Students who took any type of alternative to the regular TAKS were not eligible. As reviewed in Chapter II, most (97% to 99%) students are required to take the regular enrolled grade level (EGL) TAKS (NCLB, 2001). Only 1% to 3% of all students are eligible to take an alternate statewide assessment aligned to EGL standards.

In order to answer the research questions, data on student performance in the classroom setting and on the statewide assessment were gathered. Campus data sets and demographics were downloaded from the Texas Education Agency (TEA) Academic Excellence Indicator System (AEIS). Information on student classroom performance on EGL standards-based curriculum was gathered from performance on TAKS and on

teacher reported end-of-year classroom grades (GRADE). Disaggregated student specific TAKS data from the Spring 2007 administration were provided by each of the three participating campuses. Data were collected on the one-time administration of the STEEP3M assessment. The state certified instructional staff (e.g., teachers, educational diagnosticians, academic specialists) who worked directly or indirectly with mathematics instruction were interviewed to provide additional information to answer the research questions. Data from these sources were combined to produce a data set for the elementary schools in the study. This chapter will discuss the instrumentation and methods used for data collection.

Participants

This study was conducted using a convenience sample from Local Education Agency (LEA) campuses with third grade classrooms. In Texas, LEAs are defined as public schools and include independent school districts as well as charter schools. In 2006-07, there were 346,237 third graders who attended 4,316 elementary schools in 1,222 LEAs (TEA, 2009b, 2009c). LEAs with at least one campus with third grade students were considered for the study. Schools were selected from publicly accessible AEIS data downloaded from the TEA Web site. Each year AEIS data is based on the October snapshot date. Actual enrollment on the date of the data collection or the date of the spring TAKS administration may vary from the AEIS snapshot date data.

School Selection

Initial contact was made to a selected list of schools in the southeast Texas area within a 200-mile radius of the researcher. Fifteen elementary schools were selected for initial contact with the goal of identifying two to three participating campuses for the study. Demographics for the schools were gathered from the TEA AEIS. In selecting campuses to contact, the researcher attempted to ensure that various campus types (e.g., urban, suburban, and rural) were considered. The criteria and selection process for the LEAs, campuses, and individual participants in the study are described in this chapter. Ultimately, the student and teacher sample for this research were from three public elementary schools in southeast Texas that volunteered for the study.

Proximity Criteria. The researcher reviewed a list selected from publicly accessible information which from the TEA AEIS of 75 LEAs that were within a 200-mile radius of the researcher. The list was then narrowed down to 15-campus based on random selection. The target set for the sample was to identify 2 to 3 participating campuses, 15 educators, and 200 third grade students.

Size, Grade Level, Scheduling, and Data Criteria. The researcher determined that initial contact should be made with the campus principal, considered the instructional leader. The principal was provided with information about the study and asked to volunteer their campus third grade for participation in it, and to inform all stakeholders of such involvement, including educational staff and parents or guardians, if appropriate. The timeline for the study was provided to the principal, with the request that scheduling accommodate the need for the researcher to administer the STEEP3M within 6 to 8 weeks of the Spring 2007 TAKS administration. Additionally, principals were asked to provide the researcher with the end-of-year classroom grades and results of the Spring 2007 TAKS. All third grade students on the selected LEA campuses were eligible to be in the student sample. All staff who worked directly with instruction or assessment of third grade students were eligible to be in the instructional staff sample.

Disclosure Criteria. Principals were provided information to share with their superintendent and other stakeholders in order to obtain approval for study participation. The study proposal and the process for the Texas A&M University Institutional Review Board (IRB) were explained to the principals. Further, principals were asked to inform the researcher of any internal LEA IRB procedures or ongoing school reform initiatives. All activities to administer the STEEP3M assessment were conducted as part of the students' general curriculum and instruction; therefore, individual parental consent was not required. The principal was also asked to ensure that a campus instructional staff member be present at all times when the researcher had access to students. The process used to determine the criteria for school selection is summarized in Table 3.1.

Table 3.1

Summary of School Selection Criteria

| <i>Criterion</i> | <i>Rationale</i> |
|---------------------------|--|
| Proximity | Within 200-mile radius of researcher so that data collection for each participating campus could be efficiently and economically completed |
| Disclosure | Principal to provide information to staff, district, and parents |
| Third grade students | STEEP3M administered to third grade students |
| Size | Number of campuses to meet sample size of at least 200 students |
| Study timeline | STEEP3M administered near time of Spring 2007 TAKS mathematics test |
| TAKS scores | Campus provided and TEA downloadable TAKS data to answer questions |
| End of year course grades | End-of-year classroom grades (GRADE) to answer questions |

Summary of School Selection Criteria. Schools were selected for the study from the 75 LEA elementary campuses considered. The following procedures were used for selection:

1. From the list of all elementary schools in Texas, 75 elementary schools were chosen, based on their location.
2. From the initial sampling frame of 75 schools, 15 elementary schools were randomly selected for contact to identify a new sampling frame of at least 2 schools for the study.
3. The selection process resulted in the selection of 3 elementary schools for participation in the study. These schools:
 - a. Agreed to inform stakeholders, provide access during timeline, and a commitment to volunteering.

- b. Were considered Academically Acceptable or Recognized by TEA.
- c. Would provide end-of-year classroom grades (GRADE) and TAKS results.

Other Criteria Considered. Since the study was to focus on the usefulness of a CBM instrument and not on student performance, TEA information on LEA and campus ratings were considered for exclusionary purposes for participation in the study. Where known, LEA campuses with academically unacceptable ratings, monitoring activities, or other regulatory issues were eliminated from consideration. Although these LEA campuses might benefit from the use of an appropriate universal screener, the researcher determined that other intervention activities ongoing in the LEAs might impact the study. A further consideration was that the study might impact ongoing improvement or intervention activities at the campus or student level.

Criteria Considered but not Used. When the criteria for participation were being established, some consideration was given to only having campuses that had received an Exemplary rating from the TEA AEIS system. An Exemplary rating indicates that all students and all student groups in all grade levels taking TAKS met a 90% standard for each subject. These criteria would have significantly limited the number of campuses and LEAs that could be considered. In addition, since on Exemplary campuses most students perform well on TAKS, most would in turn perform well on STEEP3M, which might skew data needed to answer the research questions. For the purposes of this study, campuses with Recognized or Academically Acceptable ratings were also considered for participation. Recognized or Academically Acceptable are TEA ratings indicating that all students and all student groups taking the TAKS test met a standard that is set by subject. The Academically Acceptable standard varies by subject, but for mathematics indicates that at least 40% of the tested students pass the TAKS test. The standard for the Recognized rating from TEA in mathematics is set at 70%, indicating that at least 70% of the tested students passed the TAKS test (TEA, 2007). Campuses with Unacceptable ratings were not considered, for reasons stated previously.

Recruitment of LEA Participants

The researcher made initial contact to the LEA campus principal on the 15 campuses through phone or e-mail. Two principals did not respond to multiple contact attempts. Four responded that their LEA Internal Review Board (IRB) could not review the request within the timeline set for the study. An additional three were eliminated because they were already using STEEP or other district-wide commercial universal screening programs. Three campuses were eliminated because of scheduling conflicts. In all, three LEAs were identified as meeting the criteria of having campuses with third graders, proximity to the researcher, appropriate size, scheduling availability, and willingness to participate in the study.

Local Education Agency 1. The first LEA, herein identified as LEA1, is a school district approximately 30 miles northeast of a large city in southeast Texas, serving students in a non-incorporated area covering 54 square miles. This suburban and rural community is primarily residential, with farming and small businesses. In 2006-07, the enrollment of the school district as reported by TEA was 3,045 students on five campuses, from early childhood through twelfth grade. Third graders in LEA1 attend school with fourth and fifth graders on one campus. All district third graders (n=227) attend one campus (TEA, 2009a).

The demographics of the LEA1 campus are listed in the summary of all LEAs in Table 3.2 for the school years 2004-07. The campus has consistently been composed primarily of students ethnically identified as white (85.2% in 2006-07). [Note: the TEA allows students and their families to self-identify their ethnicity and self-report other personal information, but does not verify the ethnicity of the student.] Over one-fourth

Table 3.2

Student Composition Percentages for 2004-2007

| | 2004-2005 | | | 2005-2006 | | | 2006-2007 | | |
|----------------------------|-----------|------|------|-----------|------|------|-----------|------|------|
| | LEA1 | LEA2 | LEA3 | LEA1 | LEA2 | LEA3 | LEA1 | LEA2 | LEA3 |
| African American | 1.2 | 2.1 | 91.0 | 1.6 | 2.5 | 87.3 | 1.5 | 1.5 | 86.7 |
| Hispanic | 8.9 | 92.5 | 8.7 | 10.9 | 93.4 | 12.5 | 12.3 | 94.7 | 13.2 |
| White | 89.1 | 5.4 | 0.1 | 86.7 | 4.1 | 0.0 | 85.2 | 3.4 | 0.0 |
| Economically Disadvantaged | 27.0 | 93.6 | 50.4 | 28.5 | 94.6 | 91.2 | 27.1 | 94.3 | 81.5 |
| Mobility | 15.6 | 22.0 | 10.8 | 15.1 | 18.4 | 15.1 | 14.8 | 26.7 | 11.5 |
| Limited English Proficient | 2.8 | 18.5 | 8.7 | 3.6 | 20.9 | 6.5 | 4.2 | 21.7 | 8.0 |

Note: AEIS numbers reported in percentages (TEA, 2009a).

(27.1%) of the students are identified as economically disadvantaged. According to TEA campus comparison statistics for 2006-07, students enrolled in this campus have a 14.8% mobility rate and 4.2% limited English proficiency. As defined by TEA, a student is considered to be mobile if he or she has been in membership at the school for less than 83% of the school year, indicating they have missed 6 or more weeks at a particular school. In 2006-07, the average number of students per teachers was 17.2. Nearly one-fourth (23.7%) of the students on this campus were considered, under TEA guidelines for 2006-07, to be at-risk for dropping out prior to graduation, based on state-defined criteria. These criteria include, but are not limited to, that the student is under 21 years of age, was not advanced from one grade to the next for one or more school years, did not perform satisfactorily on a readiness test or assessment instrument administered during the school year, has been placed in an alternative education program, is a student of limited English proficiency, or is homeless.

As seen in the summary Table 3.3 for all LEAs in the study, most third grade students who attend LEA1 have consistently mastered the mathematics portion of the TAKS for the school years 2004-05 (89%) and 2005-06 (89%) (TEA, 2009a). In 2006,

Table 3.3

Third Graders Meeting TAKS Standard for Math 2004-2006

| | 2004-2005 | 2005-2006 |
|------|-----------|-----------|
| LEA1 | 89.0 | 89.0 |
| LEA2 | 62.0 | 66.0 |
| LEA3 | 94.0 | 88.0 |

Note: AEIS numbers reported in percentages (TEA, 2009a).

LEA1 was rated Academically Acceptable by TEA. Table 3.4 provides the student enrollment count for 2005-07.

Table 3.4

Student Enrollment Count Data Multi-Year History for 2005-2007

| | 2005-2006 | | | 2006-2007 | | |
|-------------|-----------|------|------|-----------|------|------|
| | LEA1 | LEA2 | LEA3 | LEA1 | LEA2 | LEA3 |
| LEA | 3040 | 9653 | 1180 | 3045 | 9786 | 1450 |
| Campus | 632 | 484 | 694 | 650 | 470 | 772 |
| Third Grade | 208 | 82 | 63 | 227 | 81 | 64 |

Note: AEIS numbers reported in percentages (TEA, 2009a).

Local Education Agency 2. The second LEA, herein known as LEA2, is a school district encompassing 21 square miles in the south and southwest portion of a large urban and suburban area in southeast Texas. The enrollment of the school district was 9,786 students in 2006-07. The LEA has 10 elementary schools, 4 middle schools, 2 high schools, and 2 alternative campuses. Third grade students attend elementary school. Only one of the 10 elementary schools participated in the study (TEA, 2009a).

There were 470 students enrolled in the LEA2 participating campus in 2006-07, 64 of whom were in the third grade, as shown in Table 3.4. In 2006-07, most of students who attended were identified as Hispanic (94.7%) and classified as economically disadvantaged (94.3%). Table 3.2 provides information on the demographics of the campus for the 2004-05, 2005-06, and 2006-07 school years (TEA, 2009a). In the 2006-07 school year, approximately one-fifth (21.7%) of the students were identified as having limited English proficiency, and over one-fourth (26.7%) were mobile. As shown in Table 3.3, approximately two-thirds of the students in third grade mastered the mathematics portion of the TAKS in 2004-05 (62%) and 2005-06 (66%). LEA2 is considered accredited by the Texas Education Agency and received an Academically Acceptable rating in 2005-2006.

Local Education Agency 3. The third LEA, herein known as LEA3, is also a public school and consists of three elementary campuses. The LEA primarily serves students who live in a large urban setting in southeast Texas. The school started as a single-campus private elementary in the 1980s. In the late 1990s, a charter proposal was approved by TEA to make the school a public school designed to serve economically disadvantaged students who were low performing and at-risk of dropping out. LEA3 serves students in kindergarten through grade 5. Currently there are three elementary campuses in LEA3, which serve prekindergarten through fifth grade. Enrollment was 1,450 for the three campuses in 2006-07 (TEA, 2009a). Only one of the three campuses participated in the study.

As shown in Table 3.4, there were 772 students enrolled in the participating campus during 2006-07, 81 of whom were in the third grade. The demographics are listed in Table 3.2. The majority of the students who attend are African American (86.7%). The LEA has served an increasing number of students who are economically disadvantaged; numbers jumped from 50.4% in 2005-06 to 81.5% in 2006-07. In 2006-07, approximately one-tenth (11.5%) of the students were classified as mobile and 8% were identified as having limited English proficiency (TEA, 2009a).

Summary of LEAs Participating in the Study. Summary information for enrollment in schools participating in the study is provided in Table 3.4. Each of the LEAs has had consistent enrollment for the past 3 years, with LEA3 showing a slight increase. The student composition percentages for ethnicity, economically disadvantaged, and Limited English proficiency have likewise remained consistent over the past 3 school years, as summarized in Table 3.2. LEA3 has seen a percentage increase in students who are economically disadvantaged. According to Table 3.3, most third grade students at the participating campuses had consistently mastered the mathematics portion of the TAKS for the years 2004-05 (94%) and 2005-06 (88%). In 2006, LEA3 received a Recognized rating (TEA, 2009a).

Individual Participants

Three categories of individual participants from the three LEA campuses were sought for participation in the study. This section will discuss them: campus administrators, instructional staff, and students.

Campus Administration Participants. Campus administrators from the three participating schools determined which students and instructional staff would participate in the study. They also arranged the date and time for the administration of the STEEP3M and Teacher Survey Interview Questionnaire (T-SIQ). Campus administrators arranged scheduling and access to students in order for the STEEP3M to be administered in the general education classroom setting. To the knowledge of the researcher, all students present on the day of the STEEP3M administration were included in the study.

Campus administrators provided information from the Spring 2007 TAKS mathematics performance and teacher assigned end-of-course mathematics grades. All personally identifiable or confidential information was masked, as guaranteed by the researcher. Campuses and individual participants were assigned an identifying code. The researcher requested, and the campus administrator made available, campus staff to be present at all times when the researcher was in proximity of student participants. Further, responsibility remained with the campus administrator to answer any questions parents

might have about the study. As previously mentioned, the researcher provided information reviewed and approved by the Texas A&M Institutional Review Board.

Instructional Staff Participants. Information was gathered from instructional staff who worked directly with instruction or assessment of students. Data were gathered from instructional staff using open-ended and closed questions from the Teacher Survey Interview Questionnaire (T-SIQ; see Appendix B for sample questionnaire) to gain information on the use of the curriculum based measurement, selection and use of instructional materials, and their initial impression of the STEEP3M. The campus administrators selected their instructional staff participants. All personally identifiable or confidential information was masked and teachers were assigned an identifying code. Table 3.5 summarizes the instructional staff eligible to participate in the study, totaling 22: 12 members from LEA1, 6 from LEA2, and 4 from LEA3. This exceeded the goal of identifying 15 educational staff.

Student Participants. The student participants were selected from third grade public elementary school students in southeast Texas. All students were eligible to participate in the study if they were enrolled in the third grade at participating campuses and took the regular state standards assessment (TAKS). Since TAKS is an assessment of student progress in the general education curriculum, it was noted during the design phase of the study that one of its limitations might be that collected data on students who qualify for additional supports might be limited, since these students may take assessments aligned to alternate standards. These additional supports might include but not be limited to special education and related services, English as a second language support, and/or special reading programs. The current study focused primarily on general education rather than general and special education.

Table 3.5

Summary of Participants Eligible for Study

| | LEA1 | LEA2 | LEA3 | Row Total |
|---------------------|------|------|------|-----------|
| Campuses | 1 | 1 | 1 | 3 |
| Administrators | 1 | 1 | 1 | 3 |
| Instructional Staff | 12 | 6 | 4 | 22 |
| Students | 227 | 64 | 81 | 372 |

Summary of Individual Participants. Campus administrators determined which students would participate, provided information and data on them, and made them available for the study. A total of 372 students on the three campuses were eligible. The total number of students identified exceeded the projected sample size of 200 originally set for the study. The researcher, a graduate student from the Texas A&M University of Educational Psychology Department, travelled to school campuses to collect data from a one-time, whole-class, group administration of the STEEP3M. All personally identifiable or confidential student information was masked and students were assigned an identifying code. Campus administrators arranged scheduling so that campus instructional staff was present at all times when the researcher was in proximity of student participants. Table 3.5 provides a summary of the number of participants eligible. There were a total of 372 students eligible for the study: 227 from LEA1, 64 from LEA2, and 81 from LEA3. There were a total of 22 eligible instructional staff members.

Instrumentation

Two data collections instruments were needed for the study:

1. *System to Enhance Educational Performance Grade 3 Focal Mathematics Assessment Instrument (STEPP3M)*
2. *Teacher Survey Interview Questionnaire (T-SIQ)*

The following section describes these instruments.

System to Enhance Educational Performance

System to Enhance Educational Performance (STEEP) is a company that develops educational tools and provides training for schools on the correct usage of these tools. Commercial assessment materials are available to conduct screening, give benchmark assessments, and monitor progress. Products include assessment kits comprised of materials for assessing skills such as oral reading fluency, maze (reading comprehension), and mathematics. The company uses a standard protocol approach to quickly identify the type of intervention needed in reading or mathematics for students who don't achieve benchmarks (STEEP, 2007b).

Third Grade Screening Measure. This research examined a copyrighted third grade universal CBM screening vehicle developed by the commercial group, STEEP, the *System to Enhance Educational Performance Grade 3 Focal Mathematics Assessment Instrument* or *STEEP3M* (STEEP, 2007a). The senior scientist for STEEP, Joe Witt, Ph.D., reported that the STEEP3M screener is based on the NCTM focal points. The STEEP3M CBM is also reported to be based on the mathematics focal points as developed by the National Council of Teachers of Mathematics (NCTM). The STEEP3M is designed to be administered to whole groups in the third grade classroom setting, with students being instructed to work individually. The STEEP3M is reported by its developers to be an evidenced based process designed to systematically evaluate the achievement of all students (STEEP, 2007b). The face and content validity of the instrument were reported by STEEP to have been previously reviewed but the results of the review were not provided to the researcher.

At the time of the administration of the STEEP3M for this study, the company was planning to use the tool in a separate grant-funded research project, and therefore it provided access to the assessment in order to gain pilot information. The STEEP3M consisted of 41 problems in eight sections over three pages. In addition to mathematics skills, reading and writing skills were required to complete the assessment, which might be a limitation of the study. The administration and answer key documents were printed for the researcher to utilize.

The graduate student researcher who is conducting this dissertation study is not participating in or benefiting from the STEEP study. No monetary or other compensation was or will be received by this researcher. This is an unfunded study. STEEP has granted the researcher permission to use copyrighted materials for the purpose of this study only. All copyrighted materials remain the property of STEEP, were and will be kept confidential, and will be used only for the stated purpose of this study. All confidential, personally identifiable information was masked.

The STEEP3M is a newly developed assessment tool based on the NCTM focal points. STEEP is also in the process of developing similar assessments for other grade levels. STEEP also previously developed products including an RtI package that contains materials for both reading and mathematics, training components, assessment tools for students, Web seminars, Web-based data management tools, and computer based interventions for mathematics and reading. At the time of the study in 2007, the cost for the entire RtI package was approximately \$10,000 for campuses with up to 500 students. Other materials and manuals for assessment may be purchased at a cost ranging from \$10 (mathematics computational fluency assessments, package of 50) to \$200 (school-wide assessment organizer) (STEEP, 2007b).

Administration Instructions for STEEP3M. The instructions for administration, the STEEP3M instrument, and the answer key were provided to the researcher in electronic format (Adobe systems portable document format [PDF]), black and white version. Likewise, the STEEP3M was printed in black and white for administration to students. The researcher used the STEEP3M standard set of instructions for administration of the assessment. The only oral instructions that the researcher provided to students were the scripted instructions. Students were observed to ensure that they followed instructions and completed the task.

Administration of STEEP3M. The researcher administered the STEEP3M during May 2007. Administration of the assessment was conducted in the whole-group general education classroom setting with start and finish times recorded. The researcher completed all scoring and data input, hand-scoring the probes and entering the data into

the Microsoft Excel program. Data were transferred into the NCSS statistical analysis program. Teacher responses were recorded electronically using the Microsoft Excel program and then coded for analysis.

Administration Time of STEEP3M. The assessment developers recommend 15 minutes for administration. The researcher determined that extending the time for the study to 20 minutes would give students more opportunity to complete test items. The purpose for extending the time was to gather more data to answer the research questions. The STEEP3M probes were collected at the end of 20 minutes even if students had not completed the assessment. The plan for administration was to test all students during a one-day onsite visit to each campus.

Features of STEEP3M Instrument

The STEEP3M, developed in 2007, consists of 41 problems arranged by focal point, divided into eight sections on three pages. Approximately one inch of blank space was provided above and below each section for student use. For the purposes of this study, administration of the STEEP3M was primarily to make determinations about the instrument and not for screening participating student abilities. The STEEP 3M is a copyrighted document and provided to the researcher for use only in this study. The following section provides a description of the instrument.

Section 1: Fill in Missing Numbers. The first section consisted of the first four items and required the student to provide the missing numbers to complete a stimulus question. The title to the section, “Fill in the Missing Numbers,” provided the only directions. No additional directions or sample items were provided on the student version of the form or through oral administration directions. For each item, a short stimulus was written on STEEP3M. For example, on one item the student was to “write the next even number” in the sequence, “60, 62, ____.” The correct answer is 64. For the STEEP3M, reading test items aloud to the student was not allowed. [Note: reading to a student is an allowable accommodation on the TAKS (TEA, 2009c).]

Section 2: Complete the Pattern. Items 5 through 10 on the instrument required the student to “fill in the missing numbers to complete the pattern.” Again, the title

provided the only directions, with no sample items provided. In this section, problems all had number patterns with blanks for missing items, such as: “10, 5, 20, 10, 5, 20, 10, 5, 20, ____.” The student was to write the number 10 in the blank to correctly complete the pattern. No clarification was provided on the test or by the researcher for terms such as “pattern.”

Section 3: Write the Number in Standard Form. Items 11 through 16 required the student to “write the number in standard form in the blank.” As with previous sections, the title provided the directions for the section, with no further clarification or definition of terminology such as “standard form.” In this section, the student was provided items such as “8 hundreds + 9 tens + 7 ones” and was to write the answer as 897.

Section 4: Fractions of Pie Charts. In this section, students viewed three pie charts, items 17 through 19, shaded in black and gray. The student was to look at the pie chart and determine “what part of the pie is shaded gray.” The section title, “Write the Correct Fraction,” provided the directions. Although the answer key provided answers in the fractions’ reduced format, the directions did not stipulate that the student was to reduce the fractions. Therefore, answers that were not reduced were accepted. For example, one pie chart had two sections shaded gray and two sections shaded black. An answer of $\frac{1}{2}$ or $\frac{2}{4}$ was accepted.

Section 5: Fraction Comparisons. Items 20 through 28 required the student to select and circle the correct answer to a problem comparing fractions. In addition to the title, each problem had a word stimulus directing the student to circle either the larger or smaller of the two fractions. Students were asked questions such as “Which is smaller, $\frac{2}{4}$ or $\frac{3}{4}$?” to which they would answer $\frac{2}{4}$. A limited amount of blank space above and below the section was provided for students to complete work to answer the problems.

Section 6: Using Data. In this section, students read the title and a stimulus instructing them to use the data in a table on voting to answer three questions, items 29 through 31. A data table was provided listing the 7 days of the week and the number of

votes. Students had to read three questions and provide the days of the week to answer the questions.

Section 7: Measurement. In this section with two problems, items 32 and 33, the student was to look at two rulers and answer “measurement” questions. For both questions, the student was required to provide an estimate of the length shown on the ruler.

Section 8: Multiplication and Division. In the last section for items 34 through 41, the student was required to fill in the blank with a number for each question. Eight multiplication and division problems were given, with one number left out. The student was to provide the missing number. For example, one item was “ $5 \times 5 = 25$; $___ \div 5 = 5$,” to which the student answered 25.

Teacher Survey Interview Questionnaire (T-SIQ)

In addition to data collected from the one-time administration of the STEEP3M, the researcher determined that it would be beneficial to get information from educational staff working directly with the assessment or instruction of third grade mathematics. It was necessary for many of these staff questions to be open-ended because of the descriptive nature of the information sought. Asking open-ended questions allowed for follow-up questions when necessary. The staff interview questionnaire was conducted through interviews rather than a self-administered one. The researcher developed a questionnaire titled *Teacher/Educator Questionnaire on Mathematics Curriculum Based Measurement*, herein known as *Teacher Survey Interview Questionnaire (T-SIQ)*. A copy of the T-SIQ is included in Appendix A.

Description of T-SIQ Instrument. The T-SIQ was developed as a semi-structured interview questionnaire using a predetermined set of open-ended questions (Merriam, 2001). According to Lindlof and Taylor (2002), the interviewer should determine well in advance the themes or topics for exploration during the interview. The questionnaire developed serves as a guide during a semi-structured interview to help researchers focus the interview on the identified topics without constraining them to a particular more structured format. The T-SIQ consisted of a short description of the study, approval by

the Texas A&M IRB, and seven questions developed based on topics that the researcher determined would provide information to answer the research questions. The questions were asked in a face-to-face interview, which allowed for flexibility and follow-up questions.

Table 3.6 provides information on the topics listed as indicators and themes for the T-SIQ. The first question was developed to gather information on assessments currently used to determine the mathematics skills of the students. The next question sought information on how teachers made decisions on teaching mathematics. The third and fourth questions asked for information on textbooks and commercial mathematics materials used by the campus. The fifth question sought information on initial perception

Table 3.6

Indicators Used in Developing T-SIQ Questionnaire

| Indicator | Theme |
|---|---|
| Current method of identifying student's level of math achievement | Methods (including CBM) to determine student's current math achievement |
| Current decision-making process for math instruction | Methods used to determine lesson plan content for math instruction |
| Current textbook/instructional materials | Methods/materials used to determine student's achievement level or determine lesson plan content for math instruction |
| STEEP3M | Initial response to/perception of STEEP3M |
| TAKS | Time commitment for standards-based assessment |
| Released TAKS | Time commitment for and use of released statewide assessment |

of the STEEP3M. This question had two follow-up questions designed to query the participants about their opinion on the use of the STEEP3M and the usefulness of the student performance data. The T-SIQ sixth and seventh questions gathered information on TAKS. Question six asked how long it took to administer the TAKS. Question seven focused on the use of released TAKS tests.

Data Sources

Five sources of data were needed to collect data to answer the research questions, as summarized in Table 3.7.

Table 3.7

Summary of Measurements for the Study

| Nomenclature | Measurements |
|--------------|---|
| AEIS | Academic Excellence Indicator System |
| GRADE | Teacher Reported End of Year Classroom Grades |
| STEEP3M | System to Enhance Educational Performance Grade 3 Focal Points Mathematics Assessment Instrument |
| TAKS | Texas Assessment of Knowledge and Skills |
| T-SIQ | Teacher Survey Interview Questionnaire |

Academic Excellence Indicator System (AEIS)

The Texas Education Agency (TEA) AEIS reports were downloaded from the TEA Web site to gather data on campus and student demographics and additional TAKS performance. Comparative information was gathered for the years 2002-07. After review, the researcher determined to use only the most current data from school years 2004-07, where available. At the time of the administration of the STEEP3M, the 2006-07 AEIS information was not available; it became available after the study's completion and before the research was written up.

Classroom Grades (GRADE)

Teacher reported end-of-year classroom grades on student mathematics performance on third grade level curriculum were collected (hence known as GRADE). Campus administrators provided data on classroom grades during May and June of 2007.

System to Enhance Educational Performance 3M (STEEP3M)

The STEEP3M was previously described under the Instrumentation section. Data on overall performance and on each item was gathered from the one-time May 2007 administration to the student participants in the sample.

Texas Assessment of Knowledge and Skills (TAKS)

Data were collected on student mathematics performance on the TAKS administered in Spring 2007. Campus administrators provided campus level data reports for mathematics. Information from the campus reports were reviewed to collect data on scaled score, standard met, and results by objectives. TAKS is currently the assessment utilized in Texas to determine progress towards mastery of state standards. As of this writing in 2009, the Spring 2007 TAKS tests have not yet been released and therefore are not available for study.

Teacher Survey Interview Questionnaire (T-SIQ)

Instructional staff was queried using a bank of open-ended and closed questions (T-SIQ) as previously described in the Instrumentation section. Data from the T-SIQ provided information to answer the research questions. T-SIQ questions gathered information on how the campus assesses the mathematics skills of students and how teachers make decisions on what to teach. Teachers and instructional staff were asked to report on textbooks or other commercial materials that the campus uses to teach mathematics and how these are utilized. Information was also gathered on teachers' initial impressions of the STEEP3M, TAKS administration time, and the use of released TAKS versions. The T-SIQ was administered during May and June of 2007.

Summary of Data Sources in Study. Table 3.7 provides a summary list of the five data sources used for the study. The T-SIQ provided qualitative data from instructional

staff. The AEIS provided publicly accessible demographic data and TAKS passage statistics. The other three, GRADE, TAKS, and STEEP3M all yielded quantitative data.

Design

This correlation research design was used to analyze quantitative data using parametric and nonparametric methods to study the STEEP3M, a commercially developed CBM. Questions were developed to gain qualitative information from instructional staff on mathematics teaching and perceptions of the CBM.

The study was conducted between April and September of 2007 on students who enrolled in third grade in the fall of 2006. Table 3.8 reviews the timeline for the study.

Table 3.8

Timeline for Study

| | |
|-----------------------|---|
| Fall, 2006 | Students enroll in third grade at LEA1, LEA2, and LEA3 |
| April, 2007 | Administration of the Grade 3 mathematics TAKS Initial contact to campus administrators |
| May, 2007 | Administration of the STEEP3M |
| May – June, 2007 | Administration of the T-SIQ End-of-year classroom mathematics GRADE assigned by teachers |
| May – September, 2007 | Gather TAKS data Gather GRADE Gather TEA data, including AEIS |

The spring administration of the Grade 3 mathematics TAKS test was conducted and the researcher made initial campus contact in Spring 2007. The STEEP3M was administered to students on 3 days in May of 2007. The researcher interviewed the teachers in May and June of 2007 using the T-SIQ. Campuses assigned end-of-year classroom grades (GRADE) in May and June of 2007. The researcher gathered data on TAKS, GRADE, and AEIS from May through of September, 2007.

Data Analysis

Descriptive statistics are used to summarize the sample's data numerically and/or graphically. Inferential statistics are used to identify patterns in the data, account for variances, and draw inferences about the larger population (all third grade students in public schools). Inferential statistics are utilized to model relationships (regression) or describe associations (correlation). Although correlation does not imply causation, two variables that tend to vary together may be causally connected, directly or indirectly.

The researcher administered the STEEP3M to whole-class groups of third grade students. Information was gathered from performance on the Spring 2007 administration of the TAKS and end-of-course classroom grades (GRADE) as assigned by teachers. Descriptive statistics are used to summarize the data from the measurement sources. There were 337 students, 22 teachers, and 3 administrators who participated in the study.

Answering the Research Questions

Research Question One. How well does the mathematics curriculum based measurement (CBM) assessment tool, the STEEP3M, reflect end-of-year student performance on state standards-based curriculum as measured by the end-of-year high-stakes TAKS test (internal consistency and criterion-related concurrent validity)? (Psychometric measurement question.)

Data analysis of this question focuses on the psychometric measurement aspects of the study. This question was answered from three data sources. The first was on student performance as evidenced by performance on end-of-year classroom performance (GRADE). The second was performance on the state standards assessment (Texas Assessment of Knowledge and Skills, or TAKS). The third was performance on the administration of the universal screener, the STEEP3M. The TAKS and STEEP3M are interval data and the GRADE is ordinal data. GRADE and TAKS were compared with STEEP3M scores for concurrent validity. The statistics utilized for this analysis are Pearson R correlation between the predictor, STEEP3M, and two criteria, TAKS and GRADE. Table 3.9 provides a summary of the data analysis plan for Research Question One.

Table 3.9

Data Analysis Plan for Research Question One (Psychometric Features)

| Analysis Unit (N) | Data source | Statistics | Level of data |
|-------------------|---------------------------------|---|--------------------------|
| Student (337) | AEIS GRADE | Correlation between STEEP 3M and TAKS | Interval (TAKS, STEEP3M) |
| Teacher (22) | STEEP3M TAKS Scaled Score | Pearson's R on two continuous scores between STEEP3M and TAKS Spearman's Rho on two continuous scores between STEEP3M and TAKS | Ordinal (GRADE) |

Research Question Two. To what extent does the STEEP3M provide information to support a multi-tiered decision-making process in the third grade school setting, such as for grouping of students and placement? (Applied analysis – utility for multi-tiered decision-making.)

The chief requirement for an assessment tool within an RtI model is that it must be able to identify students needing more intensive instruction. Because this was a pilot administration of the STEEP3M, it could not ethically be used for that purpose. However, it could be used to match the two strongest existing criteria for students requiring more intensive help: a) failing grade in mathematics course (GRADE) and b) failure on the high-stakes end-of-year mathematics standards-based test (TAKS).

This question is answered by logistic regression, using STEEP3M scores to predict two dichotomous variables: GRADE (pass/fail) and TAKS (pass/fail). Table 3.10 provides a summary of the data analysis plan for Research Question Two.

Table 3.10

Data Analysis Plan for Research Question Two

| Analysis Unit (N) | Data source | Statistics | Level of data |
|-------------------|---|---|--------------------------|
| Student (337) | GRADE STEEP3M | Analysis of qualitative data | Interval (TAKS, STEEP3M) |
| Teacher (22) | TAKS Results by Objective TAKS Scaled Score T-SIQ | Descriptive statistics Logistic regression | Ordinal (GRADE) |

Research Question Three. To what extent does the STEEP3M permit teachers to generate detailed mathematics lessons and unit planning for whole-class and individual needs? (Classical item analysis.)

This question focuses on application to determine the utility of the STEEP3M for planning instruction. The data sources for answering this question are STEEP3M sub-test scores and teacher interview data. Table 3.11 provides a summary of the data analysis plan for Research Question Three. The STEEP3M sub-test scores were matched to the TAKS sub-skill identification of the Texas Essential Knowledge and Skills (TEKS). Additional qualitative data (T-SIQ) were collected from teachers on their perception of the STEEP3M. The T-SIQ were analyzed by qualitative response coding. The qualitative data were obtained from a semi-structured interview between the researcher and 22 instructional staff. The researcher organized the interviewees'

Table 3.11

Data Analysis Plan for Research Question Three

| Analysis Unit (N) | Data source | Statistics | Level of data |
|-------------------|--|---|--------------------------|
| Student (337) | GRADE TEKS Standards | Descriptive statistics | Interval (TAKS, STEEP3M) |
| Teacher (22) | STEEP3M TAKS Results by Objective TAKS Scaled Score TEKS Standards T-SIQ | Logistic regression Analysis of qualitative data | Ordinal (GRADE) |

responses into segments using short word phrases that suggested how the data were associated. Once organized, the data were summarized based on the prevalence of codes, similarities and differences, and comparisons (Merriam, 2001). Analyses to answer this question were descriptive, regression, correlation, and analysis of qualitative data.

Research Question Four. How efficiently can the one-time, group administered STEEP3M be administered to third grade students (as indicated by time to administer, score, report results, and provide feedback to teachers)?

To answer this research question, data were collected on time to administer, score, report results, and provide feedback to teachers. Additional qualitative data were collected from instructional staff to further answer this question. Data sources were the STEEP3M, TAKS Scaled Score, and T-SIQ. Five statistical measures were used to answer this question: descriptive statistics, logistic regression, time recording,

observations, and anecdotal information. The STEEP3M and TAKS data were interval. Table 3.12 provides an overview of the data analysis plan for Research Question Four.

Table 3.12

Data Analysis Plan for Research Question Four

| Analysis Unit (N) | Data source | Statistics | Level of data |
|----------------------|-----------------|---|-----------------------------|
| Student (337) | STEEP3M TAKS | Anecdotal information | Interval (STEEP3M, TAKS) |
| Teacher (22) | T-SIQ | Descriptive statistics Logistic regression Observations Time recording | |

Instrument Software for Data Analysis

Organization of Data. Microsoft Office Excel is a spreadsheet application written and distributed by Microsoft for Microsoft Windows. The software features calculation and graphing tools. The first version was released in 1985 and has been the most widely used spreadsheet program since 1993 (Microsoft, 2009). Excel interfaces with other Microsoft Office software programs including Microsoft Word, the word processing software. The researcher utilized both Microsoft Excel and Word for data organization.

Software for Analysis. In 1981, the Number Cruncher Statistical System (NCSS) was developed as a computer program for statistical and data analysis. Currently identified by its acronym, NCSS, this software continues to be used for analysis and display of data. The software provides over 200 documented statistical and plot procedures. NCSS imports and exports all major spreadsheet, database, and statistical

file formats including Microsoft Excel (NCSS, 2009). The researcher utilized NCSS for statistical analysis of the research questions.

Design Limitations

This research is a study of the usefulness of a brief assessment, the STEEP3M. As such, certain assumptions have been made on the content and face validity of the STEEP3M based on information provided by the developers. To the researcher's knowledge, no pilot study was conducted on the instrument.

Limitations of Sample. The sample was made of three schools that are in close proximity to the researcher and that volunteered for the study. Given the study's sample composition, size, and focus, the results may not be used to infer conclusions beyond the scope study. However, the results may expand the research on the accuracy of a brief CBM as a performance indicator on standards-based assessment, provide information regarding the validity for planning instruction, and evaluate the ability of such a CBM to aid in screening students for a multi-tiered decision-making process.

Statewide Assessment Limitations. A potential limitation of the study would be if the TEA or Texas policymakers changed the state standards (TEKS) or decided to utilize another assessment tool or method, such as end-of-course examination. Also, according to the TEA, most students complete the TAKS in 2 to 3 hours. However, the TAKS is an untimed test and students may take an extended, unlimited amount of time to complete it. Therefore, a second potential limitation might be a time limit for the screener's administration. However, the study's focus is to identify students who demonstrate accuracy and fluency in completing the assessment.

Differentiation of the Study

This research differed from previous studies in that it focused on mathematics rather than reading; conducted assessment in the whole-group setting in the general education classroom; and utilized an assessment design that was based on national standards.

CHAPTER IV

RESULTS

This purpose of this study was to examine the usefulness of a brief mathematics assessment tool, the *System to Enhance Educational Performance Grade 3 Focal Mathematics Assessment Instrument*, STEEP3M. This study involved the selection of three elementary schools for the administration of the STEEP3M assessment to third graders in the spring of 2007. These students had taken the high-stakes state standards-based assessment test, the Texas Assessment of Knowledge and Skills (TAKS), in April 2007. Four qualities of the STEEP3M were examined: a) internal consistency and criterion-related validity (concurrent); b) screening students for a multi-tiered decision-making process; c) utility for instructional planning and intervention recommendations; and d) efficiency of administration, scoring, and reporting results. These four qualities were the basis for the four research questions for the study.

Research Question One

How well does the mathematics curriculum based measurement (CBM) assessment tool, the STEEP3M, reflect end-of-year student performance on state standards-based curriculum as measured by the end-of-year high-stakes TAKS test? (Internal consistency and criterion-related concurrent validity)

Research Question Two

To what extent does the STEEP3M provide information to support a multi-tiered decision-making process in the third grade school setting, such as for grouping of students and placement? (Screening students for a multi-tiered decision-making process)

Research Question Three

To what extent does the STEEP3M permit teachers to generate detailed mathematics lessons and unit planning for whole-class and individual needs? (Utility for instructional planning and intervention recommendations)

Research Question Four

How efficiently can the one-time, group administered STEEP3M be administered (as indicated by time to administer, score, report results, and provide feedback to teachers) to third grade students? (Efficiency of administration, scoring, and reporting results)

This chapter is organized to provide summary information on the research study and the answers to the research questions.

Participants and Setting

Three elementary campus administrators in Texas volunteered their staff and third grade students for the study. The three campuses were all public local education agencies (LEAs) herein known as LEA1, LEA2, and LEA3. Two are campuses in independent school districts and one is a public charter school. As reviewed in Chapter III, the original target for the sample size was 15 instructional staff and 200 students. The actual number of instructional staff and student participants exceeded these original targets. All participants over the original planned sample size who met specific criteria were included in the study to provide additional information to answer the research questions. Table 4.1 summarizes the total number of participants from whom data were collected. Twenty-two instructional staff members participated in the study. The STEEP3M was administered in 18 elementary classrooms at 3 separate Local Education Agencies (LEAs) to a total of 355 students during May 2007.

TABLE 4.1

Summary of All Participants in Study

| | LEA1 | LEA2 | LEA3 | Row Total |
|---------------------|------|------|------|-----------|
| Campuses | 1 | 1 | 1 | 3 |
| Administrators | 1 | 1 | 1 | 3 |
| Instructional Staff | 12 | 6 | 4 | 22 |
| Students | 207 | 60 | 70 | 337 |
| Column Total | 221 | 68 | 76 | 365 |

Students in the final sample were required to have available data from three sources, including the one-time administration of the STEEP3M, GRADE, and TAKS. As shown in TABLE 4.2, 337 of the 355 students who were administered the STEEP3M also had GRADE and TAKS data. Eighteen students who took the STEEP3M were eliminated from the study because data were not available from all three sources. Of these, nine were eliminated because they took alternative statewide assessments (e.g., the TAKS-Alternative field test [n=4], or the State Developed Alternative Assessment [n=5]). The other students did not differ demographically from the remaining sample but were eliminated because they were absent or transferred into or out of the campus during the TAKS administration. As shown in Table 4.2, the total student sample, for which the researcher had complete data from all sources, was 337, with LEA1 (n=207), LEA2 (n=60), and LEA3 (n=70). Additional information on each participating LEA is provided in Chapter III.

TABLE 4.2

Summary of Student Count by Data Source

| | STEEP3M | TAKS | GRADE | Complete |
|--------------|---------|------|-------|----------|
| LEA1 | 218 | 228 | 209 | 207 |
| LEA2 | 64 | 61 | 65 | 60 |
| LEA3 | 73 | 70 | 72 | 70 |
| Column Total | 355 | 359 | 346 | 337 |

Note: Complete = Data available from all three sources.

Descriptive Information on Data Sources

To assist in answering the four questions posed for this study on the technical adequacy, internal consistency, criterion-related validity, usefulness, acceptability, and efficiency of the STEEP3M, four additional types of data were collected. First, in May and June 2007, data on teacher-reported end-of-year classroom grades (GRADE) were collected from the three participating campuses. Second, in May and June 2007, information from instructional staff was gathered through use of the Teacher Survey Interview Questionnaire (T-SIQ) developed for this study. Third, between May and September 2007, the campuses provided student performance data for the April 2007 third grade mathematics TAKS administration. Finally, between May and September 2007, data were collected from the Texas Education Agency (TEA) public access information Academic Excellence Indicator System (AEIS). Since final data from TEA are available for public access only after a verification period, the researcher reviewed all TEA information again using the AEIS and LONESTAR public access data retrieval systems between December 2008 and March 2009. The information from a single late spring administration of the STEEP3M, the nearly concurrent administration of the high-stakes TAKS test, and other data on classroom performance (GRADE) were combined with information from the T-SIQ to answer the research questions for the study. The following section provides descriptive statistics on the STEEP3M, GRADE, and TAKS.

Descriptive Information on the STEEP3M

Descriptive statistics quantitatively describe or summarize the distribution of values for a set of observations. Descriptive statistics were conducted on the obtained scores of 337 students from the one-time administration of the STEEP3M to determine central tendency, variability, spread, and shape of the distribution.

Central Tendency of STEEP3M. Measures of central tendency represent the score of a typical individual in the group (Crocker & Algina, 2008). The three most common measures of central tendency are the mean, median, and mode. The mean, or average number of items correct, was 29.43 (70.73%, standard deviation [SD] = 6.59); the mode, or most frequently occurring score, was 33; and the median, or number that lies at the midpoint of the distribution, was 31.

Variability of STEEP3M. Measures of variability indicate the degree of the dispersion among the spread of the scores. For the STEEP3M with a mean of 29.37, the range was 4 to 40. None of the 337 students answered all 41 items correctly. Whereas the range provides an index of the dispersion among the full group of scores, the interquartile range (IQR) indicates how much spread exists among the middle 50% of the scores, and is considered a more robust measure of dispersion because it is not affected by outliers (Huck, 2004, 2008; Sax, 1989; Thorndike, 1997). If a distribution is concentrated around the mean, the IQR will be small. Likewise, if the data is widely dispersed, the IQR will be large. The IQR is the difference between the 25th and 75th percentiles. For this distribution, the IQR was 26.5 (25th percentile) to 34 (75th percentile).

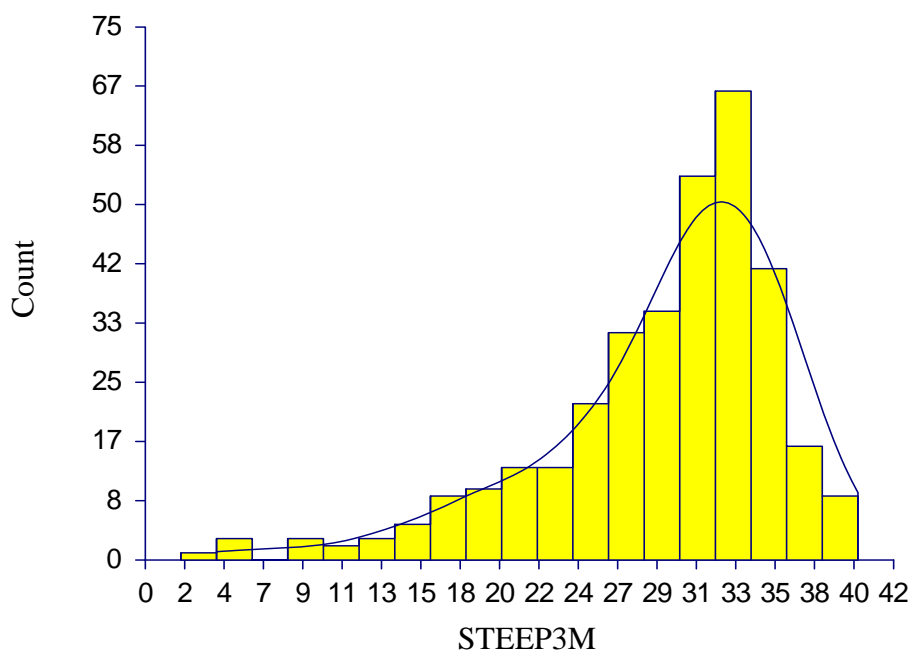
Distribution of STEEP3M. The shape of a distribution describes the pattern of the numbers along the number line. As shown in Figure 4.1, the STEEP3M scores are highly skewed to the left, or negatively skewed, indicating that the majority of the values lie above the mean value. Skewing refers to the extent to which a distribution of values deviates from symmetry around the mean. Skewing one way or another will tend to reduce the test's reliability. A negatively skewed distribution may indicate that a test was too easy, with many high scores and few low scores (Huck, 2004; Sax, 1989). According

to Sax (1989), negatively skewed distributions of items that are relatively easy are favored if the purpose is to place students in classes for struggling learners or for targeting interventions or identifying students who have not mastered the curriculum.

There is a sharp rise between 21 and 31 items answered correctly (see Figure 4.1), indicating that correctly answering relatively few items could make a big difference in a student's percentile rank. That is not an ideal result and indicates that the STEEP3M is lacking in well-graduated items at the higher difficulty range. Discrimination is the ability of an item to separate higher ability participants from lower ability ones (Thorndike, 2005). The STEEP3M does not appear to discriminate very well and a relatively few items above the 25th percentile could make a big difference in a student's percentile rank. Descriptive statistics for the STEEP3M are summarized in Table 4.3, following the descriptive statistic information for the GRADE and TAKS.

FIGURE 4.1

Histogram of Number of Items Correct for STEEP3M



Note: N=337 Students.

Descriptive Information on GRADE

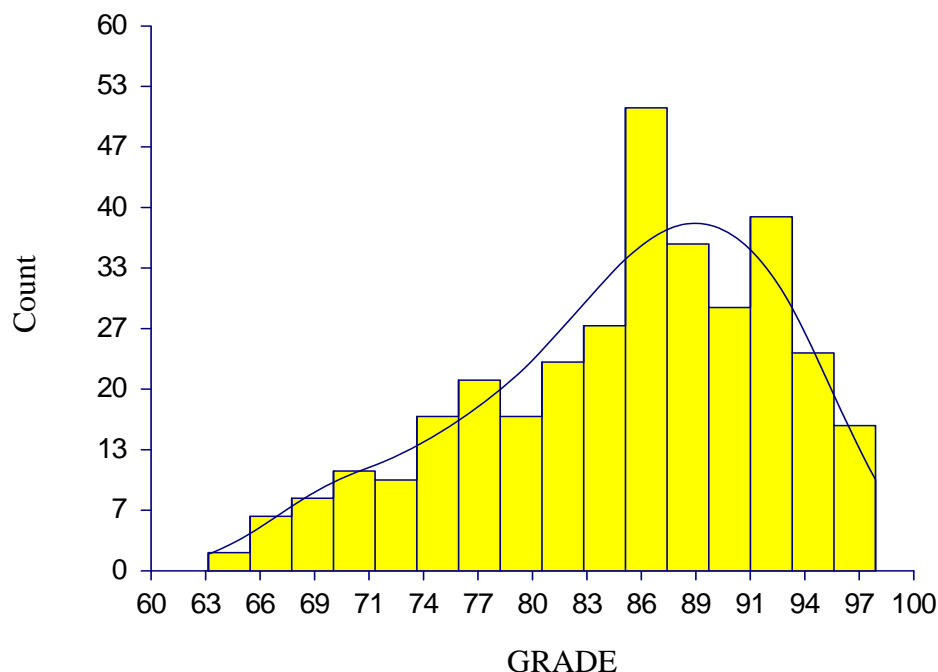
The grading scale for students in the study was based on teacher assigned classroom grades from zero to 100. Grades of 70 or above were considered to be passing, reflecting that the student has mastered the curriculum in the classroom setting.

Central Tendency of GRADE. The mean end-of-year report-card grade (GRADE) for students participating in the study was 84.84. The median was 86 and the mode was 92.

Variability of GRADE. The GRADE scores ranged from 63 to 98. The IQR was calculated on the middle 50% of the GRADE scores as a more robust statistic. The IQR of these distribution scores was 80 to 91.

Distribution of GRADE. Only 6 students (1.78%) received grades in the failing range (below 70), which may be considered to skew the data because of the subjectivity of teacher assigned grades. An unbalanced distribution such as this one will make it difficult to measure strong relationship with any other variable (Huck, 2004). Strong effects, whether from analysis of variance (ANOVA), correlation, or other type of statistical analysis, rely on well-balanced data. As shown in Figure 4.2, the shape is quite smooth, but as was the case with STEEP3M, the data were skewed negatively to the left, indicating most grades were above the mean. Descriptive statistics for GRADE are summarized in Table 4.3.

FIGURE 4.2
Histogram of Math End-of-Year Grade (GRADE)



Note: N=337 Students.

Descriptive Information on TAKS

The third grade mathematics TAKS was administered to 297,734 students in elementary schools across Texas in April 2007. The TEA set the standard for passing this TAKS at the scale score of 2100. According to the TEA (2009c), the mean scaled score was 2263. The test consisted of 40 items across 6 third grade mathematics TEKS objectives (herein known as TEKS 3-1, TEKS 3-2, TEKS 3-3, TEKS 3-4, TEKS 3-5, and TEKS 3-6).

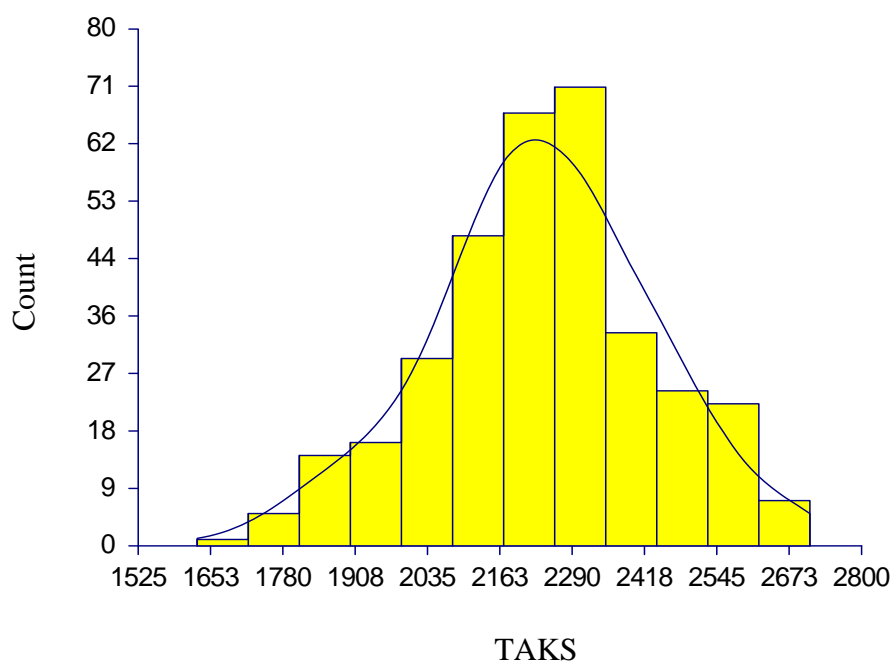
Central Tendency of TAKS. Descriptive statistic analysis of the TAKS scaled scores for students in the study indicated the mean was 2236.86, which is slightly less than the statewide mean of 2263. The median was 2232 and the mode was 2400. Although the mean, median, and mode are not equal, they are fairly close.

Variability of TAKS. The range of TAKS scaled scores for students in the study was 1629 to 2709 (mean 2236.86, SD = 194.6). The robust IQR for the TAKS was 2115 (25th percentile) to 2400 (75th percentile).

Distribution of TAKS. The TAKS distribution for students in the study is considered normal with an almost perfect bell shape around a single mean, as shown in Figure 4.3. Descriptive statistics for TAKS are summarized in Table 4.3, following the Summary of Information on Data Sources section.

FIGURE 4.3

Histogram of TAKS Scale Score for Students in Study



Note: N=337 Students.

Summary of Information on Data Sources

Data were collected on 337 students for the study from three sources: 1) a one-time administration of the STEEP3M; 2) end-of-year mathematics grades (GRADE); and 3) the April 2007 administration of the TAKS third grade mathematics tests (see

Table 4.2). Descriptive statistics for STEEP3M, GRADE, and TAKS for the students in the study are summarized in Table 4.3.

TABLE 4.3

Descriptive Statistics Summary for STEEP3M, GRADE, and TAKS

| | STEEP3M | GRADE | TAKS |
|---------------------------------|---------|-------|---------|
| Minimum | 4.00 | 63.00 | 1629.00 |
| Mean | 29.00 | 84.84 | 2236.86 |
| Median | 31.00 | 86.00 | 2232.00 |
| Mode | 33.00 | 92.00 | 2400.00 |
| Maximum | 40.00 | 98.00 | 2709.00 |
| Standard Deviation | 6.59 | 7.85 | 194.61 |
| IQR 25 th Percentile | 26.50 | 80.00 | 2115.00 |
| IQR 75 th Percentile | 34.00 | 91.00 | 2400.00 |

Note: N=337 Students; STEEP3M based on number of items correct; TAKS based on Scaled Score.

Highly skewed variables will not give excellent results because of the nature of the correlations (Huck, 2004; Sax, 1989; Thorndike, 2005). The correlations will be attenuated or dwarfed because of the lack of spread on one side of the distribution. The negative skewing of the STEEP3M data is primarily due to the large number of participants who scored high on the STEEP3M, indicating that for many it is too easy a test. Likewise the negative skewing of GRADE data indicates that a high number of students received a high end-of-year course grade, which will make it hard to get a strong relation with any other variable.

Research Question One

How well does the mathematics curriculum based measurement (CBM) assessment tool, the STEEP3M, reflect end-of-year student performance on state standards-based curriculum as measured by the end-of-year high-stakes TAKS test? (Internal consistency and criterion-related concurrent validity.)

The description of instrument quality primarily focuses on two measurement-related concepts: reliability (internal consistency) and validity. Therefore, the first research question is focused on these psychometric properties of the STEEP3M as measured by internal consistency and criterion-related concurrent validity. The STEEP3M is a criterion-referenced measurement tool purported to predict TAKS mathematics end-of-year performance. To answer this question, data were collected and analyzed from three sources, the STEEP3M, GRADE, and TAKS. The item-level database included 41 STEEP3M items and 337 students.

Internal consistency is the extent to which the parts of a measurement instrument, in this case the individual questions on the STEEP3M, measure the same thing (Huck, 2004, 2008). Other forms of reliability (e.g., retest, alternate forms) are difficult to achieve at an acceptable level if the instrument has poor internal consistency. Internal consistency may be determined from correlational analysis of a single one-time administration of a test to a single group of individuals (Crocker & Algina, 1986, 2008; Huck, 2004, 2008). Spearman's Rho and Cronbach's Alpha output were reviewed to help determine the internal consistency of the STEEP3M (Crocker & Algina, 1986, 2008; Huck, 2004, 2008).

Criterion-related validity checks a test's relationship to other, accepted, or established measures in the same area, and is a standard requirement for a new measure (Crocker & Algina, 1986, 2008; Huck, 2004, 2008; Salkind, 2004). The two main types of criterion-related validity are concurrent and predictive. Concurrent validity examines the test's relationship with a "criterion measure," both administered at close to the same time (Crocker & Algina, 1986, 2008; Huck, 2004, 2008; Sax, 1989; Thorndike, 2005). The strength of the correlation of the STEEP3M with two criteria, the TAKS (Pearson's R) and the GRADE (Spearman's Rho) were reviewed. According to Sax (1989), criterion-referenced tests present some unique problems because many items may be too easy, and domains are difficult to define even if objectives are carefully worded (Sax, 1989; Thorndike, 2005). Item analysis is used to determine the difficulty of an item and

how well an item distinguishes or discriminates between higher and lower scoring examinees (Thorndike, 2005).

In answering this question, classical item analysis was conducted to review summary statistics on item performance. The key elements of classical item analysis include a) the item difficulty of each item; b) the item discrimination of each item; and c) the overall Cronbach's Alpha (Sax, 1989; Thorndike, 2005). Distributions of item difficulty and item discrimination can show the levels of ability for which a test is best and worst suited. Efficiency and/or speediness of the test for the particular group are also a part of classical item analysis. This will be reviewed in Research Question Four.

Internal Consistency

Internal consistency analysis was used to determine the extent to which items on the STEEP3M were consistent with one another in that each item represents one, and only one, single dimension, construct, or area of interest. The researcher administered the STEEP3M to third grade students on one occasion to gather data to judge reliability. The reliability of the test was estimated by analyzing how well the STEEP3M items correlated with one another. High inter-item correlation and Cronbach's Alpha indicate that the items all measure a single construct or ability, in this case, third grade math ability. Cronbach's Alpha measures how well a set of items (or variables) measures a single, one-dimensional construct (Huck, 2004; Thorndike, 2005). For this study the Cronbach's Alpha, discussed further in this chapter, was used to assess the internal consistency of the STEEP3M.

Item Difficulty. Item difficulty is the proportion of participants who answered a given item correctly (Gulliksen, 1950, 1987; Sax, 1989; Thorndike, 2005). For test items on the STEEP3M there was one right answer for each item, and therefore the answer was dichotomous (right/wrong). Item difficulty of the STEEP3M was assessed using the item's average, or mean score. The average item mean for the STEEP3M is .72 as shown in Table 4.4, indicating the assessment tool was on the borderline of being too easy. The average item difficulty value was 80%, with an IQR of .67 to .86, which represented the percentage correct by the average student across 41 items for 337 students. The average

item difficulty at the 10th percentile was 25%, indicating that only 25% of the students got them correct. The most difficult items were at or below the 10th percentile. The item difficulties were not well distributed, with many easy items plus a few very difficult ones (6). From this table, it is apparent that there is a large jump from the 10th to the 25th percentile and then a moderate jump from the 25th to the 50th percentile. The item difficulty level then does not change much between the 50th (.800), 75th (.858), and 90th (.929) percentiles. Items that are extremely easy or hard cannot effectively discriminate among students (Sax, 1989). A test should have a fairly even distribution of item difficulties, with some increase in number of items toward the performance mean (Sax, 1989). This will prevent the necessity of large jumps in performance to increase the total score scale. For the STEEP3M, there was a large undesirable gap between the easy and difficult items. Overall, the STEEP3M was quite an easy test.

TABLE 4.4

Item Analysis for STEEP3M: Average Item Difficulty

| | 10th | 25th | 50th | 75th | 90th |
|-----------|------------|------------|------------|------------|------------|
| Parameter | Percentile | Percentile | Percentile | Percentile | Percentile |
| Value | 0.257 | 0.671 | 0.800 | 0.858 | 0.929 |

Table 4.5 reports the average item difficulty or “percent correct” of TAKS. The IQR for this distribution is .70 to .93, with a mean of .78 and a standard deviation of .175.

TABLE 4.5

Item Analysis for TAKS: Average Item Difficulty

| | 10th | 25th | 50th | 75th | 90th |
|-----------|------------|------------|------------|------------|------------|
| Parameter | Percentile | Percentile | Percentile | Percentile | Percentile |
| Value | 0.500 | 0.700 | 0.825 | 0.925 | 0.950 |

Item Discrimination. Item discrimination is a measure of how well an item is able to distinguish between examinees with high versus low overall test scores (Huck, 2004; Sax, 1989; Thorndike, 1997, 2005). The possible range of discrimination is between -1 and 1. An item that discriminates negatively may indicate that the item is measuring something other than what the rest of the test is measuring, because a higher percentage of examinees are answering correctly (Crocker & Algina, 1986; 2008; Sax, 1989; Thorndike, 1997, 2005). In 1965, Robert Ebel provided the following guidelines for item discrimination: 1) discrimination values of .40 or above are satisfactory; 2) those between .30 and .39 require little or no revision; 3) those between .20 and .29 are marginal and need rewriting; and 4) those below .19 should be discarded or significantly revised (Crocker & Algina, 2008; Payne, 2003).

As shown in Table 4.6, the STEEP3M item discrimination has a mean of .52, with a range of .23 to .82. The IQR is .35 to .67, indicating that about 10% of the items in the STEEP3M are marginal and need rewriting.

TABLE 4.6

Item Analysis for STEEP3M: Average Item Discrimination

| | 10th | 25th | 50 th | 75th | 90th |
|-----------|------------|------------|------------------|------------|------------|
| Parameter | Percentile | Percentile | Percentile | Percentile | Percentile |
| Value | 0.284 | 0.353 | 0.534 | 0.673 | 0.741 |

Note: Quartile Section of STEEP3M. Average Item R^2 (Item Discrimination) Mean = .52. Range .23 to .82.

Cronbach's Alpha. Cronbach's Alpha is the average value of the reliability coefficients that would be obtained for all possible groupings of items, split in half (Cronbach, 1951; Cronbach & Shavelson, 2004; Gliem & Gliem, 2003; Huck, 2005). If test items measure the same construct, then the two resulting subsets should correlate. The closer the correlation is between the two subsets the greater the internal consistency. George and Mallery (2003) suggest the following rules of thumb for evaluating the Cronbach's Alpha coefficients: a) greater than .90 is considered excellent; b) .80 to .89 is considered good; c) .70 to .79 is considered acceptable; d) .60 to .69 is considered questionable; e) .50 to .59 is considered poor; and f) .49 or below is considered unacceptable. Coefficients in the high 80s or 90s are common for established tests. Utilizing NCSS statistical software for analysis, the standardized Cronbach's Alpha was .88 for the STEEP3M, which is considered good and suggests the questions comprising the test are internally consistent, mainly measuring a single mathematics ability construct.

Criterion-related Validity

In psychometrics, criterion-related validity, either predictive or concurrent, is a measure of how well a target test or subtest predicts an external established (and usually well-accepted) measure termed the "criterion" (Thorndike, 1997, 2005). One of the major uses of a criterion-referenced test such as the STEEP3M is for making predictive decisions about the examinee's level of competency on a domain of skills (Allen & Yen, 1979, 2002; Crocker & Algina, 1986). Concurrent validity is established by comparing a

new test, in this case the STEEP3M, with measures that are already considered valid or widely accepted. A high positive correlation with the existing measures would be interpreted as establishing validity for the new measure. In this study, the two criterion measures for measuring concurrent validity were course grades (GRADE) and the high-stakes year-end TAKS test, which closely mirrors the TEKS mathematics curriculum standards for Texas. If criterion measures are quite similar in content and item type, and are themselves reliable, then validity coefficients of .70 to .85 are expected. If the match between the target test and criterion measure not as close, or if the criterion measure itself lacks reliability, then coefficients of only .50 to .60 are expected. This study utilized two criterion measures, the TAKS and GRADE. Although GRADE was used as a measure of subject mastery and to determine whether a student advances to the next grade level, it has unknown, probably low, validity. GRADE is an aggregate of individual assessments, which may include many subjective measures (Allen, 2005; Marzano, 2000).

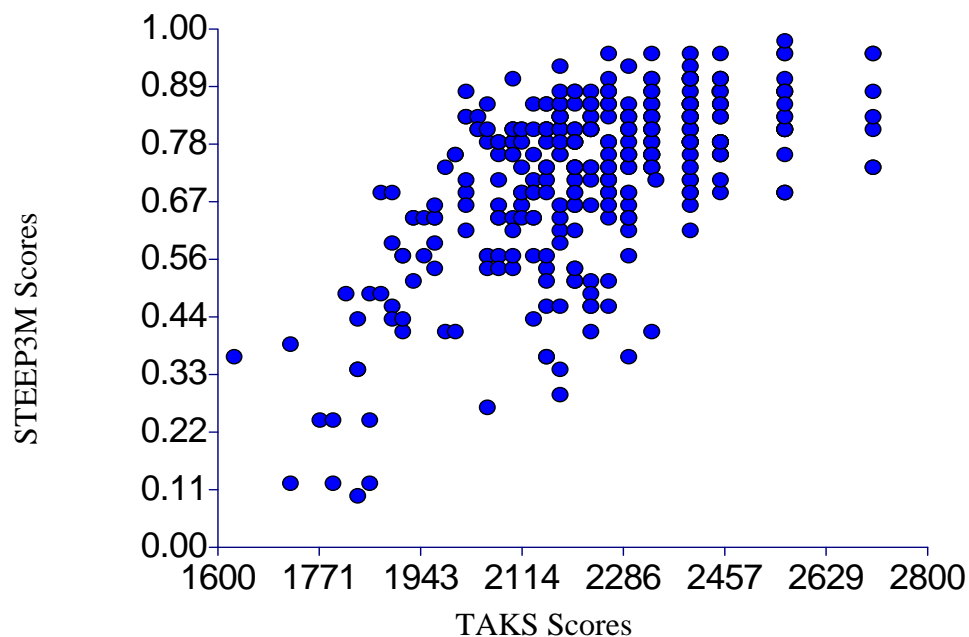
STEEP3M predicted the *TAKS* moderately well, with $R = .597$ ($p < .000$). STEEP3M's prediction of the criterion GRADE was similar, at $R = .587$ ($p < .000$). Since both validity correlations were similar, the question was whether GRADES and TAKS were basically measuring the same thing. A follow-up multiple regression was conducted, predicting STEEP3M from a combination of GRADE and TAKS. The result was $R = .65$, a little higher than either of the independent relationships. This means that GRADE and TAKS do share considerable variance and tend to measure the same thing, even though they are independent measures. The common ingredient is probably the mathematic ability of the student, plus study habits.

Differential Predictability of TAKS by STEEP3M. Differential predictability means that for some skill levels the predictor does not work as well as for other skill levels (Ghiselli, 1963; Jorgensen, 1970; Vinchur, 1993). To check differential predictability from STEEP to TAKS for various students, the student group was divided into three equal sub-groups according to TAKS scores, and STEEP3M predictability was checked within each group. The sample size ($N = 337$) was large enough to provide

three ordered sub-groups of 112 students each. Note that because the range of each of these sub-groups was restricted, lower correlations would be expected. Results by level between STEEP3M and TAKS showed strong differences. Correlations were as follows for the low, medium, and high group. For the low group, the $R = .64$ ($p < .000$), for the medium group, the $R = .0425$, and for the high group, the $R = .273$. The STEEP3M shows the poorest discrimination among TAKS levels in the middle of the range, followed by the upper range. The absolute size of each analysis by level should drop on the basis of attenuation alone, yet that is not the case for the low score range. The uneven predictability of the TAKS versus STEEP3M scores can be seen in the scatter plot in Figure 4.4. A slight diagonal tendency of the TAKS scale score up to about 2000, and then a flat tendency, reflects low correlation in the middle and upper regions of the TAKS scores.

FIGURE 4.4

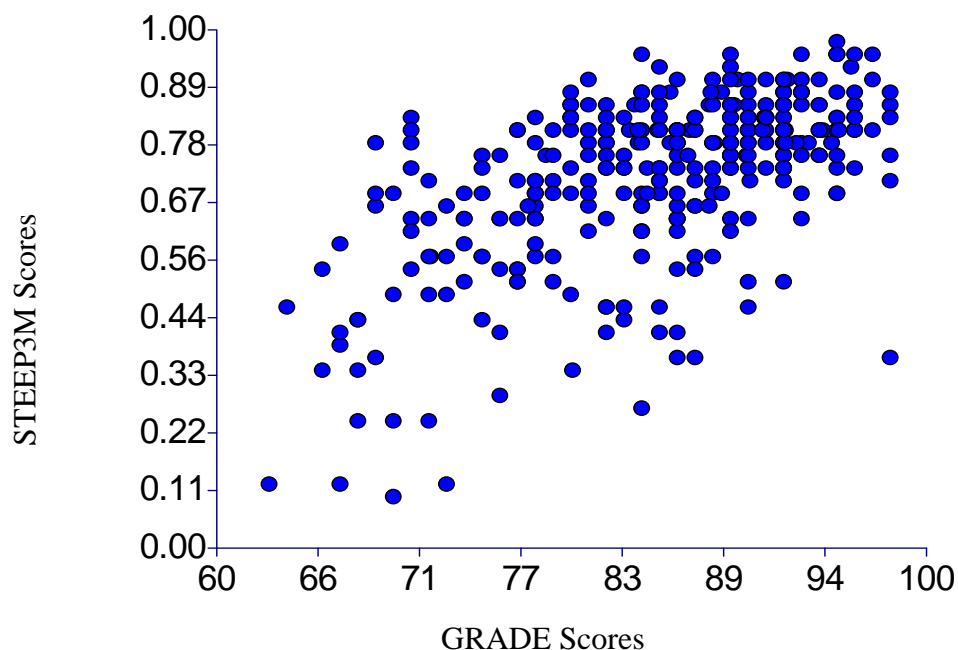
Scatter Plot of TAKS versus STEEP3M Scores



Differential Predictability of GRADE by STEEP3M. The same question on differential predictability was then asked about the STEEP3M in predicting end-of-year classroom grades (GRADE). The grade scale was divided into three segments with 112 students in each segment. As with TAKS, because the range of each subgroup was restricted, lower correlations would be expected. For the lowest group of students, $R = .536$, for the middle group, $R = .243$, and for the highest group, $R = .343$. Thus, as a predictor of GRADE, STEEP3M's effectiveness varies greatly by the student's GRADE level. Prediction is poorest for students with a middle GRADE, and next poorest for students with a high GRADE. It is a reasonably good predictor only for students with a low GRADE. In Figure 4.5, the flat GRADE section from about 80 upward represents a low correlation with STEEP3M scores.

FIGURE 4.5

Scatter Plot of GRADE versus STEEP3M Scores



Summary of Research Question One

To answer Question One on the consistency and validity of the STEEP3M, analyses of internal consistency and criterion-related concurrent validity were conducted. Classical item analysis was used to determine item difficulty, item discrimination, and the Cronbach's Alpha. Overall the STEEP3M was quite an easy test, with an average item difficulty IQR of .67 to .86. The item discrimination IQR of .35 to .67 indicated that most items had acceptable discrimination, with about 10% needing to be rewritten. The Cronbach's Alpha of .88 is considered good and suggests the test questions are internally consistent.

Regarding the second part of Research Question One on criterion-related validity, the sample of 337 students was subdivided into three groups for analysis for criterion-related concurrent validity. Based on these analyses, the STEEP3M shows poorest discrimination among TAKS levels in the middle of the range. Discrimination at upper range is also poor. It is a reasonably good predictor within the lowest third of class, by GRADE.

Research Question One addressed two important psychometric properties of the STEEP3M regarding internal consistency and concurrent validity. These two qualities provide foundational information to answer Research Question Two on the utility and Research Question Three on the usefulness of the STEEP3M for teachers. A test that did not possess internal consistency or validity would not provide information for a multi-tiered decision-making process for grouping and placement of students or for planning instruction.

Research Question Two

To what extent does the STEEP3M provide information to support a multi-tiered decision-making process in the third grade school setting, such as for grouping of students and placement?

The second research question focused on the utility of the STEEP3M for screening students within a multi-tiered decision-making process. According to information on STEEP's Web site, students whose needs are not being met by the core

curriculum can be identified by universal screening for intervention and progress monitoring. STEEP3M is purported by its developers to be a universal screener for multi-tiered Response to Intervention (RTI) models that help schools determine whether a student needs differentiated instruction, special grouping, or placement in a special program (STEEP3M, 2007). These decisions are generally made after failure has occurred in the classroom or on a standards-based assessment (TAKS). Therefore, the developers of the STEEP3M promote that it will help schools reduce inappropriate referrals to special education. Most RTI models identify three tiers of instruction, the first being the basic curriculum for whole-group instruction, for approximately 85% of the students. The second and third, or bottom, tiers (for approximately 15% and 5% of students, respectively) provide specific small group or individual targeted interventions (Fuchs & Fuchs, 2007).

To evaluate the degree of decision accuracy for a given test within an RTI framework, one “true criterion” or more is needed. For this study, the true criteria were pass/fail for TAKS and GRADES. Failure on either criterion has serious consequences. It will likely result not only in decisions on reteaching, but in-class support grouping, special programs or intervention supports placement, and/or moving a student into the next RTI level. It would be useful to have a test with an RTI model that could predict either of these pass/fail criteria. The standard method for judging accuracy in predicting pass/fail is by the Receiver Operator Curve (ROC), accessed either separately in a “diagnostic testing” module or within Logistic Regression. This analysis yields the probabilities for four outcomes for a dichotomous decision: true (or actual) positives, false positives, true negatives, and false negatives. The ROC is a graphical representation of two characteristics: the sensitivity versus the specificity for a binary classifier system (Hosmer & Lemeshow, 2000). Central to using ROC curves is a cutoff level to classify participants into one of two groups – in this case, students who are in danger of failing the third grade mathematics TAKS, and those who are not (Allen & Yen, 2002; Crocker & Algina, 1986, 2002). An ROC curve is the plot of the true

positive rate (sensitivity) against the false positive rate (specificity) for the different possible cutoff points for a diagnostic test.

The calculation for sensitivity, or the degree to which a diagnostic instrument correctly identifies children at-risk (in this case, those at-risk for failing the third grade mathematics TAKS) involves dividing the true positives (predicted to fail and actually failed) by the sum of the true positives and the false negatives (students who were predicted to pass but indeed failed). A sensitivity of 100% would mean that all students who might fail the TAKS were correctly identified. A test with low sensitivity would miss the target and would not predict those students who need help.

Specificity is the degree to which a diagnostic instrument correctly identifies students at low risk for failing the TAKS. Therefore, for this study, a specificity of 100% means that the test correctly identified all students who would pass the TAKS. A test with low specificity would falsely identify students as being at-risk of failing the TAKS who are not actually at-risk of failing. Low sensitivity or low specificity results in false alarms that, by misidentifying students, cost time and money and cause inconvenience and unneeded parent and/or student stress. Specificity is calculated by dividing the number of true negatives (students who were predicted to pass TAKS and did so) by the sum of true negatives and false positives (students who were predicted to fail TAKS but passed).

Therefore, there is a direct relationship between specificity and sensitivity: as one increases, the other decreases. A fundamental purpose of this study is to determine the usefulness of the STEEP3M as a universal screener to identify students who might be in danger of failing the TAKS, scoring failing GRADES, or requiring an RtI intervention. Increasing the likelihood of any testing errors has a major impact on high-stakes decisions for educating students. A false positive error might incorrectly place a student in remediation or intervention programs, which would limit general classroom or enrichment education instruction time. A false negative might prevent a student from receiving intervention, additional instruction, or special programs, causing them additional struggle in the general classroom setting.

Categorical decision analysis within logistic regression was used to test the ability of the STEEP3M to predict TAKS pass or failure. This dichotomy, TAKS pass or fail, is a real-world scenario with implications for educating a student differently, based on performance or predicted performance on the TAKS. Additionally, decisions on changing instructional settings, referral for special programs including special education, selection of instructional materials, student grouping, and setting goals for students are all impacted by a student's performance on the TAKS. This analysis began with an ROC analysis from logistic regression yielding a classification accuracy table for predicting whether or not the student met the TAKS standard, a scaled score of 2100.

Logistic regression is a method for determining the maximum likelihood of occurrence by using the ROC to optimize prediction of classification accuracy. The predictors in logistic regression are one or more continuous variables, and the dependent or criterion measure is a categorical (often dichotomous) placement (George & Mallery, 2003; Hosmer & Lemeshow, 2000; Huck, 2008). Logistic regression provides more flexibility than other analyses because it does not assume linear relation between variables or a normally distributed dependent variable (Huck, 2004). There are several overall optimized solutions for analysis in order to capture the most failures in prediction, as shown in Table 4.7. The cutoff score for predicting pass/fail can be changed to decrease any one of the two types of errors (false positives or false negatives), but at the cost of an increase in the other type of error.

TABLE 4.7

Classification Accuracy Table for Prediction of Whether Student Met TAKS Standard

| A | B | C | D | Sensitivity | Specificity | Classification Accuracy |
|----|-----|----|-----|-------------|-------------|-------------------------|
| 63 | 232 | 1 | 41 | .984 | .150 | .308 |
| 54 | 120 | 10 | 153 | .843 | .560 | .614 |
| 50 | 79 | 14 | 194 | .781 | .710 | .724 |
| 45 | 50 | 19 | 223 | .703 | .816 | .795 |
| 35 | 33 | 29 | 240 | .546 | .879 | .816 |
| 34 | 28 | 30 | 245 | .531 | .897 | .827 |
| 32 | 27 | 32 | 246 | .500 | .901 | .824 |
| 23 | 19 | 41 | 254 | .359 | .930 | .821 |
| 22 | 14 | 42 | 259 | .343 | .948 | .833 |
| 19 | 13 | 45 | 260 | .296 | .952 | .827 |

Note: A=Actually failed, predicted failed (true positive); B=Actually passed, predicted failed (false positive); C=Actually failed, predicted passed (false negative); D=Actually passed, predicted passed (true negative).

For the logistic regression models in this study, the researcher was most interested in classification accuracy and sensitivity, and then in observing the associated specificity. In the context of RtI, the sum of true and false positives constitutes the sample for secondary tiered interventions (Fuchs et al., 2007). Many RtI models suggest that 5% (for this study, $n=17$) of students will be in tertiary intervention and 15% (for this study, $n=51$) in secondary intervention, bringing the sum of true and false positives to 20% ($n=68$) (Fuchs et al., 2007). In the actual sample for the study, 19% ($n=64$) actually failed the TAKS. Therefore, adjusting the cutoff point for optimized solutions requires a balance; the net must be cast widely enough to catch students at-risk for failing TAKS, without falsely catching those who are not at-risk. The cost of low sensitivity is in missing the target, whereas the cost of low specificity is in identifying too many students for intervention.

To demonstrate utilizing an optimized solution, the overall ROC default classification accuracy for predicting a TAKS pass/fail rate of 79.5% was chosen from the 10 alternatives shown in Table 4.7. The true positive (n=45, or 13%) and false positives (n=50, or 15%) would equal a student group (n=95, or 28%) targeted for RtI intervention. The sensitivity of the STEEP3M at this optimized solution was 70.3% (45/64). The value for specificity was 81.6% (223/273), and considered good. Under this solution, 19 students (5%) would be predicted to pass but actually failed (false negative), and therefore would not be identified for intervention by the STEEP3M. Table 4.8 provides a visual for the selected optimal solution.

TABLE 4.8

Selected Optimized Solution

| | | Actual | | |
|-----------|--------------|--------|------|-----------|
| | | Fail | Pass | Row Total |
| Predicted | Fail | 45 | 50 | 95 |
| | Pass | 19 | 223 | 273 |
| | Column Total | 64 | 242 | 337 |

Note: Classification accuracy = 79.5%.

At the overall classification accuracy of 72.4%, the percentage of students identified for intervention would be 38% (n=129), with the misidentification of students who actually failed being approximately 2.9% (n=10). The sensitivity at this solution was 84.3% and the specificity was 71%. If the LEAs in the study had the resources to provide intervention to more than one-third of their third grade students, this solution would be the better choice, as it would miss fewer at-risk students. With this solution, however, about 50 students who were not in danger of failing TAKS would receive additional support. In contrast, choosing an overall classification accuracy of 81.6% would decrease the percentage of students targeted for intervention to 20% (n=68);

however, 8.6% (n=29) of the students who actually failed would not have been identified for intervention. With this solution, specificity was 87.9% and sensitivity was 54.6%.

Teacher Perception of the STEEP3M for RtI

The T-SIQ was administered in interview format to 22 instructional staff who work with students in mathematics, including teachers, educational diagnosticians, and instructional specialists. The majority of the instructional staff (n=20, or 90%) indicated that they would definitely consider using the STEEP3M for RtI multi-tiered decision-making to identify students for targeted interventions, with the other 2 indicating they would possibly use it. In response to how they would use the STEEP3M, their answers included: as a screener at the beginning of the year (n=12, or 54%), to identify students in need of tutoring and remediation (n=9, or 41%), and to group students in the classroom (n=8, or 32%). Although teachers indicated they might use the STEEP3M, they stated that it would not be the only criteria for placing students in more intensive interventions for third grade mathematics. Teachers reported that they would use it as a part of a rubric for decision-making, combined with the previous year's mathematics TAKS scores and classroom grade, as well as other curriculum based measurement tools.

Summary for Research Question Two

Categorical decision analysis and qualitative analysis of teacher data were conducted to answer Research Question Two on the extent to which the STEEP3M can be used as a universal screener to make initial decisions on student grouping and placement in a multi-tiered intervention process. Logistic regression and ROC analysis was conducted in answering this question. By adjusting the cutoff score, several alternative solutions were identified. Of these optimized solutions, three were identified for further review. Although each had their attractions, none provided a solution that came without significant cost to the school such as misidentification of students at-risk of failing or creating large groups for multi-tiered intervention. Two of the optimized solutions reviewed would have created groupings for RtI of approximately one-third of the class: 28% (n=95) and 38% (n=129) respectively. At a classification accuracy of

79.5%, the true positives (n=45, or 13%) and true negatives (n=50, or 15%) would equal a student group of 28% (n=95) targeted for RtI intervention. Under this solution, 5% (n=19) of students would be predicted to pass, but actually failed (false negative) and therefore would not be identified for intervention by the STEEP3M.

Teachers' initial perceptions of the STEEP3M indicated they would consider using it as a part of a decision-making rubric rather than totally depending on the results of a one-time administration to identify struggling students. Information from categorical analysis and qualitative analysis of teacher data both indicate that although the STEEP3M might provide useful information to support a multi-tiered decision-making process in the third grade school setting, such as for student grouping or placement in mathematics interventions, it would be insufficient to identify the majority of the students who might be struggling.

The next question will review the extent to which the STEEP3M can support decisions for student lesson planning. Predictive correlations between subscales of the STEEP3M and TAKS will be conducted to answer Research Question Three, in order to provide a more complete answer to the overall usefulness of the STEEP3M to predict performance on the TAKS.

Research Question Three

To what extent does the STEEP3M permit teachers to generate detailed mathematics lessons and unit planning for whole-class and individual needs?

Research Question Three focused on the STEEP3M sub-skills and the extent to which they match the standards-based mathematics curriculum. To answer this question, three specific areas were addressed for quantitative analysis, and one for qualitative analysis. First, a content analysis of TEKS and STEEP3M was conducted to determine how well the STEEP3M covered the third grade curriculum. Second, classical item analysis determined how well the STEEP3M represented TEKS from a measurement point of view. Third, in order to address how adequately STEEP3M sub-groups related to the TAKS, using TAKS as the defining criterion for good coverage of TEKS, correlation analysis was conducted between STEEP3M and the standards-based

curriculum (TEKS/TAKS). Fourth, data from the T-SIQ were analyzed to determine the face validity of the STEEP3M. Information between STEEP3M sub-groups and TAKS will also provide additional information to answer Research Question One. To answer Research Question Three, data and information were gathered from six sources: a) GRADE (ordinal); b) TEKS (state standards curriculum; c) the one-time administration of the STEEP3M (interval); d) TAKS scale score (interval); e) TAKS results by objective (interval); and f) T-SIQ. The researcher's observations on the instrument, including information on the alignment of the STEEP3M to the TEKS, are described in the following sections.

Content Analysis

The examination of the statistical properties of the STEEP3M content-based sub-tests was conducted to help determine whether teachers could use the results to support planning instruction. Analysis of content validation assesses whether the items on a test represent the domain or construct (Crocker & Algina, 2008), in this case, the third grade state standards-based mathematics curriculum.

In Texas, the standards-based curriculum, the Texas Essential Knowledge and Skills (TEKS), is reported by TEA to be aligned to the National Council of Teachers of Mathematics (NCTM) focal points (TEA, 2009c). The developer and senior researcher of the STEEP3M, Joe Witt, Ph.D., reported to this researcher that the STEEP3M was aligned to the NCTM focal points and therefore would be aligned to the TEKS, as measured by the TAKS.

STEEP3M Content Alignment to Curriculum. The third grade TAKS is a 40-item multiple choice test that measures a student's performance on six objectives of the TEKS (TAC, 2006). The first objective (TEKS 3-1) is numbers, operations, and quantitative reasoning. According to TEA (2009c) there are 10 questions on the TAKS that align to the TEKS 3-1. The second objective (TEKS 3-2) is patterns, relationships, and algebraic reasoning, with 6 questions on the TAKS. The third objective (TEKS 3-3) is geometry and spatial reasoning, with 6 questions on the TAKS. The fourth objective (TEKS 3-4) is concepts and use of measurement, with 6 questions on the TAKS. The fifth objective

(TEKS 3-5) is probability and statistics, with 4 questions on the TAKS. The sixth objective (TEKS 3-6) is mathematical processes and tools, with 8 questions on the TAKS.

As shown in Table 4.9, the STEEP3M does have items that measure each of the six TEKS objectives represented on the TAKS. Of the 41 items on the STEEP3M, the majority of questions (23 items, or 56%) align to TEKS 3-1 which focuses on number, operations, and quantitative reasoning. For analysis purpose, the items in this sub-group were broken down into three groups within the TEKS objective. The first grouping aligned to TEKS 3-1 had 6 items (14.6%) that examined how well the student used place value to communicate about increasingly large whole numbers. The second grouping of 9 items (22%) examined how well the student compared fractions and also aligned to TEKS 3-1. The third grouping aligned to TEKS 3-1 consisted of 8 items (19.5%) and examined how well the student solved and recorded multiplication and division problems. Approximately one fourth of the STEEP3M items (10, or 24%) measure TEKS 3-2, which examines how well the student understands patterns and relationships.

Approximately 80% (33/41) of the items on the STEEP3M measure TEKS 3-1 and 3-2. The remaining items (as shown on Table 4.9) measure TEKS 3-3 (7%), TEKS 3-4 (5%), TEKS 3-5 (2%), and TEKS 3-6 (5%). This means that some of the STEEP3M sub-test groupings have very few items, and therefore may not perform well in terms of internal consistency. This information will be very important to teachers as these numbers may be too few for diagnostic use, especially since the questions are multiple choice or short answer rather than production items. The four groups measuring TEKS 3-1 and 3-2 have at least 6 items, and therefore may be a reasonable measure of mathematics sub-skill weakness. A limitation of the STEEP3M is that it lacks items that measure time, temperature, or money. This will be further discussed in Chapter V.

TABLE 4.9

Summary of Number of STEEP3M Test Items to TEKS Objectives

| TEKS Objective | STEEP3M | Cronbach's Alpha |
|---|---------|------------------|
| 3-1 Numbers, Operations, & Quantitative Reasoning: | | |
| Place Value | 6 | .874 |
| Fractions | 9 | .766 |
| Quantitative Reasoning | 8 | .740 |
| 3-2 Patterns, Relationships, & Algebraic Reasoning: | 10 | .705 |
| Patterns | | |
| 3-3 Geometry & Spatial Reasoning | 3 | .680 |
| 3-4 Concepts & Use of Measurement | 2 | .442 |
| 3-5 Probability & Statistics | 1 | n/a |
| 3-6 Mathematical Processes & Tools | 2 | .464 |

Classical Item Analysis

If sub-item groupings are to be useful in detailing instructional strengths and weaknesses, each sub-test score must be composed of items which are similar in content and which have reasonably good inter-item correlations. Classical item analysis was conducted on eight STEEP3M sub-test item groupings by TEKS objectives: a) TEKS 3-1 Place Value (items 11-16); b) TEKS 3-1: Fractions (items 20-28); c) TEKS 3-1: Quantitative (items 34-41); d) TEKS 3-2: Patterns (items 1-4; 5-10); e) TEKS 3-3: Spatial Reasoning (items 17-19); f) TEKS 3-4: Measurement (items 32-33); g) TEKS 3-5: Statistics (item 29); and h) TEKS 3-6: Processes (items 30-31). The relevant score was Cronbach's Alpha. These results must be interpreted cautiously because Cronbach's Alpha is known to be partially dependent on the number of items in the sub-scale

(Duhachek, Coughlan, & Tacobucci, 2005). As shown in Table 4.9, the Alphas are from good (.874) to moderate to fairly weak (.442).

Descriptive statistics analyses were conducted on the STEEP3M sub-groupings. As shown in Table 4.10, the sub-groupings show very poor discrimination meaning that for some of these groupings, performance at the 90th and the 10th percentiles are nearly the same, or differ by only a few items. Instructional staff would not find this information useful for diagnostics. Since this information might be based on one or two

TABLE 4.10

Descriptive Statistics on STEEP3M Sub-groupings by TEKS Objective

| TEKS | Mean | SD | Percentile | | | | |
|-----------------------|------|------|------------------|------------------|------------------|------------------|------------------|
| | | | 10 th | 25 th | 50 th | 75 th | 90 th |
| 3-1 Place Value | .753 | .330 | .167 | .667 | .833 | 1.000 | 1.000 |
| 3-1 Fractions | .646 | .228 | .222 | .667 | .667 | .778 | .800 |
| 3-1 Quantitative | .853 | .271 | .375 | .875 | 1.000 | 1.000 | 1.000 |
| 3-2 Patterns | .829 | .187 | .580 | .700 | .900 | 1.000 | 1.000 |
| 3-3 Spatial Reasoning | .437 | .341 | .000 | .333 | .333 | .667 | 1.000 |
| 3-4 Measurement | .280 | .359 | .000 | .000 | .000 | .500 | 1.000 |
| 3-5 Statistics | .546 | .499 | .000 | .000 | 1.000 | 1.000 | 1.000 |
| 3-6 Processes | .786 | .330 | .500 | .500 | 1.000 | 1.000 | 1.000 |

Note. N=337 Students; SD=Standard Deviation.

items it would not be considered trustworthy. A student might get one item right or wrong, but the score would not be a good predictor of whether or not the skill was mastered. Peculiarities of the individual item performance, including answers obtained by chance, make it difficult to generalize to decision-making for identifying content for lesson planning, choosing teaching methods, and/or grouping/placement decisions.

Correlation Analysis of Sub-groups

A total test score does not provide information on which specific skills a student may or may not have; therefore, results were reviewed on individual items or sub-groups of items. Sub-groups of items tend to have greater reliability than individual items. Studying sub-groups cancels out some of the individual item anomalies and allows discussion on level of mastery of content sub-areas and sub-domains.

To continue answering Research Question Three on how well the STEEP3M could provide information to support a multi-tiered decision-making process, the researcher conducted classical item analysis. Content analysis would not provide information on how well the STEEP3M represented TEKS from a measurement point of view. Classical item analysis involves the basic measurement concepts of determining the test score, the true score, and the error score. Further, classical item analysis utilizes item- and sample-dependent statistics such as item difficulty and item discrimination to determine correlations and reliability of scores (e.g., Cronbach's Alpha), as well as test difficulty (Schumacker, 2005). Classical item analysis is useful when constructing or piloting tests because it can be conducted with smaller samples of test takers and utilizes analysis of item difficulty and item discrimination.

The results of the classical item analysis for TEKS 3-1 on place value indicated that the six items on the STEEP3M that reflected TEKS 3-1 differed in mean from .588 to .828, with an average of .75. Average item-item correlations were .68. The Cronbach's Alpha was .869 and considered good. One item had a correlation of .474 and the other five items in this sub-grouping had correlations ranging from .622 to .780.

The next sub-group analyzed was also aligned with TEKS 3-1, a student's ability to compare fractions. The Cronbach's Alpha for this sub-group was .76 and considered acceptable. The means ranged from .139 to .893, indicating much within-group variance of items. Two items within this group had a mean of less than .2 and five had means of over .8. Two items within this sub-grouping had correlations of .053 and .091, and six others had correlations between .562 and .623.

The last sub-group of items aligned to TEKS 3-1 measured quantitative reasoning. This sub-group showed a fair amount of consistency in the calculations, with smaller amounts of variance. The means were between .733 and .917, and Cronbach's Alphas were between .885 and .918.

The Cronbach's Alpha for the sub-group of 10 STEEP3M items aligned to TEKS 3-2, Patterns, is .705 and considered acceptable. The means were between .650 and .970, with correlations between .164 and .516.

Likewise, according to the data as shown, the Cronbach's Alpha (.680) is questionable for the sub-group aligned to TEKS 3-3, Spatial Reasoning. This might be attributed to a small number of items (3) in the sub-group. The means ranged from .252 to .774, with correlations between .245 and .667.

The results for the classical item analysis for the STEEP3M grouping aligned to TEKS 3-4, Measurement, indicates the Cronbach's Alpha is .442 and in the unacceptable range. Again, this might be attributed to the small number of items in the sub-group. Since TEKS 3-5 had only one item aligned on the STEEP3M, analysis could not be conducted. As with the two previous subgroups, the Cronbach's Alpha (.464) for the subgroup of STEEP3M items aligned to the TEKS 3-6, Processes, is in the unacceptable range.

Correlation between STEEP3M and TAKS. The majority of correlations between the STEEP3M and TAKS were weak, ranging between .098 and .531 as shown in Table 4.11. Correlations that fall between 0 and 1.00 are considered positive or direct. Correlations between .10 and .30 are considered small, between .30 and .50 are considered medium, and between .50 and 1.00 are considered large (Huck, 2008). Correlations over .50, in the large range, would support the diagnostic usefulness of the STEEP3M for lesson planning, at least for TAKS preparation in the area of mathematics. Correlations in the range of .30 or below are generally not useful and considered weak. The correlation between the STEEP3M sub-group on place value and TEKS 3-1 is the only one in the acceptable range (.531). A weak correlation can be due to several things. First, inadequate sampling could be a cause, which would not be the case here. There

were 337 students in the sample from diverse backgrounds. Also, a lack of internal consistency in STEEP3M or TAKS, lack of fit in content between the two tests, or very different item types or formats between the two tests could cause a weak correlation. Research Question One provided information that the STEEP3M has acceptable internal consistency. This will be examined more closely in Chapter V. Table 4.11 shows the correlations between the STEEP3M and TAKS by TEKS objective sub-group alignment.

TABLE 4.11

Correlation between STEEP3M and TAKS by TEKS Objective

| TAKS by TEKS Objectives: The student will demonstrate an understanding of: | STEEP Correlation | Number STEEP Items |
|--|-------------------|--------------------|
| 1 Numbers, Operations, and Quantitative Reasoning: Place Value | .531 | 6 |
| 1 Numbers, Operations, and Quantitative Reasoning: Fractions | .297 | 9 |
| 1 Numbers, Operations, and Quantitative Reasoning: Quantitative | .199 | 8 |
| 2 Patterns, Relationships and Algebraic thinking | .453 | 10 |
| 3 Geometry and Spatial Reasoning | .321 | 3 |
| 4 Concepts and Uses Of Measurement | .098 | 2 |
| 5 Probability and Statistics | .192 | 1 |
| 6 Processes | .301 | 2 |

As a comparison, according to Pearson (2009), the developers of the KeyMath3, a norm-referenced measure of essential mathematical concepts and skills, report a split half internal consistency for kindergarten through Grade 5 as: a) basic concepts as .92; b) operations as .89; c) applications as .85; and d) total test as .95. Although the KeyMath3 is an individually administered test, takes longer to administer and score, and does not purport to be aligned to the criterion of state standards, it does provide a comparison for reliability for measures available for instructional staff use.

Face Validity

Face validity refers to how well the test appears to measure a construct from the point of view of an experienced end-user (Crocker & Algina, 2008). For the purpose of

this study, information on the lay perspective was gathered from educators using the T-SIQ. Data from the T-SIQ were coded and analyzed to determine the face validity by asking typical users to examine the STEEP3M and its parts. Analysis of face validity helps to determine whether the questions are relevant to the intended target content and comprehensively cover it. The T-SIQ interview was structured to collect qualitative data on teacher perception of the assessment tool's utility, methods campuses use to assess mathematics skills, how decisions are made on what to teach, and instructional materials. Information from instructional staff was also collected on the perceived impression of the instrument as well as the data's usefulness.

Information from Instructional Staff. Information was gathered through the T-SIQ on methods and data used to make decisions for lesson planning. Under the No Child Left Behind Act of 2001, instructional staff members are the primary assessors of student performance toward mastery of the standards-based curriculum. These federal mandates carry the expectation that instructional staff have the ability to measure a student's current skill level, determine needs for interventions, select and implement interventions, deliver instruction, and administer the standards-based TAKS test. To validate the test, input from instructional staff on the utility of the STEEP3M for instructional planning and intervention recommendations is needed.

Teacher Reported Methods for Decision-making. Respondents indicated that their LEAs utilize benchmarking, TAKS disaggregated data, weekly assignments, mathematics facts tests (a timed test with 100 problems), and verbal questions to assess mathematics skills for their RtI process. Three (3/22, or 14%) reported that even though their LEA provides benchmarking, classroom teachers assess skills by working with students or verbally questioning them. Each teacher indicated that these methods were time-consuming, taking hours from classroom instruction during the first weeks of the school year. Although all teachers indicated they received disaggregated data from the previous year's TAKS, this provided little predictive information on how the students would perform on the current year's state standards as assessed by the TAKS. Teachers

indicated that performance on the previous year's TAKS was not used to group students for RtI interventions.

Most (15/22, or 68%) of the respondents stated that teams of teachers meet in the summer or before school starts to review and agree on a scope and sequence of curriculum objectives to teach. Of these 68%, all indicated that they follow this sequence and do not have much input on the selection of instructional materials or on changing sequencing based on student need. Some (3/22, or 14%) indicated that they meet weekly to review student progress and plan lessons based on student needs. Meeting more frequently allows instruction to be more responsive to the student. According to the teachers interviewed, this scope and sequence was neither based on data from assessment of students' needs nor on the previous year's TAKS scores. All indicated that the plan developed in the summer might not match the needs of the incoming third grade students, and that performance data from a screener of student skills at the beginning of the year was not used to adjust lesson plans.

Mathematics Textbook and Commercial Materials. Of the respondents, most (17/22, or 77%) said they use *Saxon Math* as a textbook. A few (4/22, or 18%) indicated that they use *Math in My World* from McGraw Hill. One (5%) respondent indicated that she rarely uses the textbook, but does use manipulatives, picture books, student-derived problems, or other resources. All (22/22, or 100%) respondents reported that they use some type of commercial mathematics materials or programs. In addition to materials that accompany textbooks mentioned above, the following materials or programs were mentioned: *Buckle Down Texas Mathematics* (2), *Destination Math* (3), *Lone Star Learning* (10), *Mentoring Minds* (1), *River Deep* (3), *Step up to TAKS* (1), *TAKS Buster* (2), *TAKS Master* (2), and *Target the Question* (9). Respondents indicated they use these materials as daily warm-ups, for vocabulary and problem solving, as supplemental materials, and for after-school tutorials. Only *River Deep* was reported to target individual student needs and is used as an intervention tool for students who are struggling.

None of the instructional staff interviewed reported that the textbook or commercial materials used had a universal screener to group students based on predicting how a student would perform on the GRADE or TAKS. All reported that it would be beneficial to have an assessment that provided predictive information for planning tiered interventions targeted to improve student performance on the TAKS.

Initial Perception of STEEP3M. As reported in Research Question Two, when asked their initial impression of the STEEP3M, the majority (20/22, or 90%) indicated that they would consider using the test to plan instruction and identify student mastery and need, and for planning small group instruction and targeted intervention. When asked whether data from this CBM would help them plan instruction, most (17/22, or 77%) responded “yes,” and 4 (19%) said that it “could” or “possibly.” If they responded “yes” to this question they were asked how the data could help plan instruction. Three of this group of 17 (18%) responded that it would help identify Tier 1 and 2 students for targeted interventions or small group instruction. Three (3/17, or 18%) indicated that it could be used for a beginning of the year screener for all students. One teacher stated that currently they have to wait until TAKS results are received to see how the campus and student will measure on the state assessment, and then it is too late to change instruction or improve student performance. Five (5/17, or 29%) stated that this data would help guide them in concepts needed for whole group or small group instruction. This assessment would help them know what each student understands and check for whole-class understanding of materials. In turn this would help plan reteaching, review lesson planning, and create classroom or grade level groups. Additionally, 5 respondents (29%) reported that they have to wait until the end of the school year or the summer to receive TAKS results, which does not allow an opportunity to adjust instruction, change lesson plans, or target grouping or interventions.

Summary of Teacher Perceptions of STEEP3M. In summary, instructional staff primarily report that they assess student skills individually in the classroom setting by verbally questioning and testing. These methods are time-consuming; taking many hours from classroom instruction during the first weeks of school. Most educators reported that

they meet in the summer to plan the scope and sequence of the curriculum for lesson planning. Further, they report that these lesson plans are followed throughout the year, with changes based solely on students' needs as identified at the beginning of the school year. Most of the instructional staff report that they use commercial textbooks for teaching even though they lack universal screeners and do not primarily target individual student needs. Even though the majority (77%) of the instructional staff reported that they would consider using the STEEP3M to help plan instruction, only a few (18%) indicated that it would be helpful to identify students for grouping or targeted interventions.

Summary of Research Question Three

In answering Research Question Three on the extent to which the STEEP3M permitted teachers to generate detailed mathematics lessons and unit planning for whole class and individual needs, four types of analysis were conducted on the STEEP3M: 1) content analysis; 2) classical item analysis; 3) correlation analysis of sub-groups; and 4) face validity. To determine the usefulness of the STEEP3M as a formative assessment to guide instruction, qualitative information from the T-SIQ was analyzed and combined with quantitative data from the STEEP3M and TAKS. Tests used for formative evaluation can guide instructional staff in planning or modifying instruction before students take summative or high-stakes tests. Various instructional decisions must be made in order to support students individually or in small groups, or in the whole classroom setting. These include targeting instruction; selecting curriculum for how to present the state standards-based curriculum; and placing or classifying students in whole class, small group, or targeted interventions (Thorndike, 1997, 2005).

Content analysis of TEKS and STEEP3M was conducted to determine how well the STEEP3M covered the third grade curriculum. To some extent, the 41 items in the STEEP3M were aligned to the 6 TEKS objectives on the TAKS. The majority (56%) of the STEEP3M items were aligned to TEKS 3-1, which focuses on numbers, operations, and quantitative reasoning. Further, most (80%) of the items on the STEEP3M were aligned to TEKS 3-1 and TEKS 3-2 (fractions). Some of the STEEP3M sub-test item

groupings have very few items and therefore may not be internally consistent. The numbers may be too few for diagnostic usefulness. Some of the questions are multiple choice or short answer rather than production items, further confounding the diagnostic usefulness. Additionally, the STEEP3M does not measure some specific skills in the TEKS such as time, temperature, and money.

Classical item analysis was done to determine how well the STEEP3M represented TEKS from a measurement point of view. The Cronbach's Alphas ranged from good (.874) to moderate to fairly weak (.442). Descriptive statistics analyses were conducted on the STEEP3M sub-groupings indicating very poor discrimination meaning that for some of these groupings, performance at the 90th and the 10th percentiles are nearly the same, or differ by only a few items. Instructional staff would not find this information diagnostically useful.

Correlation analysis was conducted to determine whether the STEEP3M sub-groups related to the TAKS, using TAKS as the defining criterion for good coverage of TEKS. Overall, there was a weak correlation between the STEEP3M and the TAKS. Although weak correlations can be attributed to several things, including small sampling and lack of student diversity, neither was the case in this study. Lack of internal consistency in the STEEP3M and TAKS might be due to lack of fit in content between the two tests.

Face validity was measured by instructional staff responses to the T-SIQ to determine the layperson's perspective on the test's ability to measure a meaningful construct, in this study the third grade mathematics curriculum. Most of the instructional staff (77%) reported that they would consider using the information from the STEEP3M to plan instruction. In contrast, only a few (18%) reported that the STEEP3M results would be used to identify students for grouping or targeting interventions.

Research Question Four

How efficiently can the one-time, group administered STEEP3M be administered to third grade students (as indicated by time to administer, score, report results, and provide feedback to teachers)?

The developers of the STEEP3M highlight the fact that it is a universal screener that can be efficiently administered and scored. For any assessment, the target is to acquire the greatest number of meaningful responses in the shortest amount of time. Gathering data takes time and comes at some expense, so it is important to use the most efficient and cost-effective method (Payne, 2003). Therefore, in addition to determining validity and reliability, it is important to consider the efficiency and cost effectiveness of administering a universal screener. Second, the relationship of the time a student uses to take the test and the score received is an indicator of their proficiency level, and can provide information for where to set cut-off scores to differentiate groups (Crocker & Algina, 2008). To answer this question, data were collected on time of administration, student completion, student performance, and scoring of the STEEP3M. Since this was a pilot study on the instrument, feedback information on actual results were not reported to the participating instructional staff or campuses. Research Question Four focused on the efficiency of the administration, scoring, and analysis of student performance in relation to STEEP3M completion time.

Administration of the STEEP3M

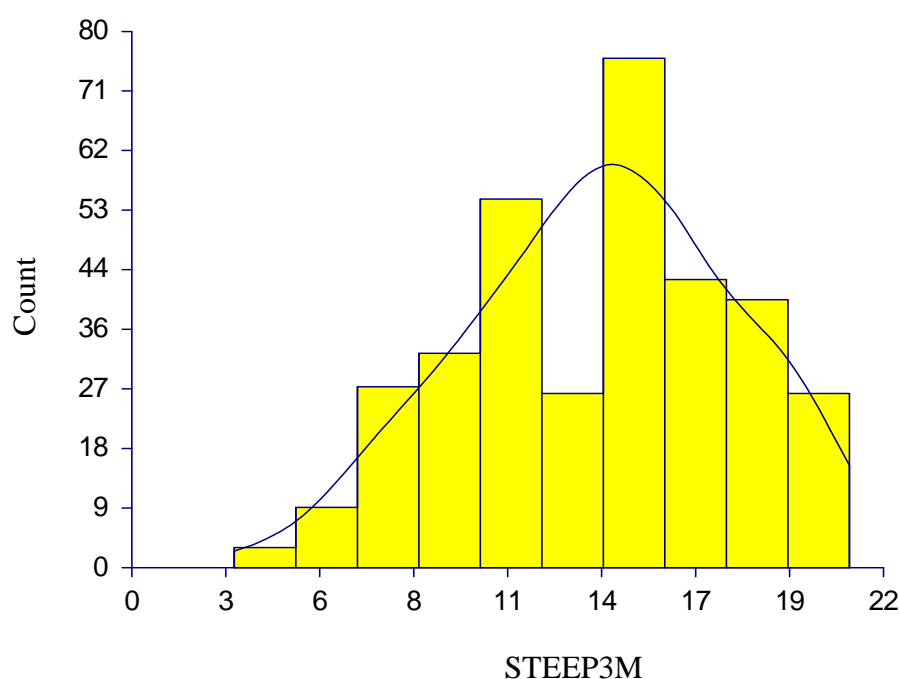
The STEEP3M was administered in a whole group setting during a one-time per campus administration with students working independently. The STEEP3M was administered to 18 classes on 3 separate campuses. The total time for delivering instructions, administering the tests, and collecting answer sheets was approximately 30 minutes per class. From the researcher's observations, all students were engaged during the administration time with minimal interruption for questioning attempts.

Time Limits of Administration. The time allotted for the assessment was 20 minutes to answer 41 questions. Students were allowed to turn in tests upon finishing. At the end of 20 minutes, all tests were collected. Fewer than 2% of the students in the study (6/337, or 1.8%) were still working on the STEEP3M when the tests were collected. Of the 337 students, 2 (.05%) finished in 3 minutes, 1 (.02%) finished in 4 minutes, and 4 (1.18%) finished in 5 minutes. Seventy-one (21%) students finished in 10 minutes or less. The average time to take the assessment was 13.69 minutes. The mode,

or most frequent amount of time, was 15 minutes. The range for time taken to complete the STEEP3M was 3 to 21 minutes. Eleven (3.3%) students did not complete all of the assessment items by the end of 20 minutes. Figure 4.6 provides a visual representation of the time for 337 students to finish the STEEP3M, which was close to a normal distribution.

FIGURE 4.6

Plots Section of STEEP3M Time Finished in Minutes



Note: N=337 Students.

Although dissemination of the information from scoring from this pilot study was not conducted, the researcher estimated that it would take approximately 10 minutes per student. In total, the per student time to administer, score, and disseminate information would be about 30 minutes.

In comparison, according to Pearson (2009) the KeyMath3, an individually administered mathematics achievement test, takes 30 to 90 minutes per student. To

individually administer the KeyMath3 to 337 students would take approximately 21 days of 8 hours each, based on the minimum administration time of 30 minutes. Other individual achievement tests such as those developed by Woodcock-Johnson and Wechsler also are individually administered and would take similar administration times. In addition, since these are individually administered tests, parental consent for evaluation would be required (IDEA, 2004).

Information from Instructional Staff

All (22/22, or 100%) of the respondents indicated that the TAKS administration generally takes 2 to 5 hours to administer. One reported that a student completed the TAKS in 40 minutes. Time for taking the TAKS is not limited, and 5 (5/22, or 23%) reported that the administration took all day, approximately 8 hours, with only a break for lunch. One stated, “If we had more than one day, it would take more than one day.”

Use of Released TAKS. Slightly more than half (12/22, or 54%) of the respondents reported that they use released TAKS third grade mathematics tests. Five (5/22, or 23%) indicated that the released TAKS is used as benchmark testing to determine the student’s mastery of third grade TEKS mathematics objectives. In addition, the student’s performance on the released TAKS is reviewed to assess mastery of skill objectives based on missed items. Additionally, 5 (5/22, or 23%) reported using it as a mock test to practice administration procedures for the “real” TAKS and to give students the opportunity to see what test day will be like. According to respondents, administration of the released TAKS to the whole group takes 1.5 to 4 hours. Two (2/22, or 9%) use portions of the released TAKS for small group activities or as daily warm-up activities. Less than half (9/22, or 40%) of the instructional staff indicated that they liked the fact that the STEEP3M was brief and could be administered in 20 minutes or less.

Relationship of STEEP3M Score and Time to Complete

The relationship between the time to take the STEEP3M and the score on the STEEP3M was negative, with $R = .157$. This indicates that the longer a student took, the fewer questions they answered correctly. Students who know less find the test more difficult, spend more time thinking and pondering, make false starts, and work through

the tests more slowly. In addition, as reviewed in Chapter II, low readers tend to also be low in mathematics skills. Low readers also are slow readers and tend to take longer on mathematics tests.

Scoring. The researcher served as the administrator and scorer for all 337 STEEP3M tests for the study. The researcher determined accuracy for intra-rater reliability with three separate scoring reviews. Scoring all 41 items on each test took approximately 2 to 3 minutes. Initial scoring of all tests took approximately 16.85 hours. In all, initial scoring and review took approximately 50.55 hours, or just over 6 eight-hour days. Scores for each item as well as for the entire STEEP3M test were hand-entered into an Excel spreadsheet and then transferred to a statistical program for analysis. Data entry took approximately the same amount of time as scoring.

Summary of Research Question Four

Research Question Four focused on the efficiency of the STEEP3M as measured by two factors: 1) time to administer and score the test, and 2) relationship of a student's time to take the STEEP3M and their score. Even if a test is reliable, internally consistent, and has validity, it will not be used in the classroom setting unless it can be efficiently administered and scored. Several other measures (e.g., released TAKS, KeyMath3) can provide specific diagnostic information on individual students; however, they take a minimum of 30 minutes to administer. This is a significant amount of classroom time that must be taken from teaching to dedicate to assessment.

The average amount of time to take the STEEP3M was less than 14 minutes. The total time for administration for an entire class was 30 minutes or less. Each test took approximately 3 minutes to score. Although dissemination of the information from scoring from this pilot study was not conducted, the researcher estimates that it would take less than 10 minutes per student. In total, the per student time to administer, score, and disseminate information would be about 30 minutes. This compares to 2 to 4 hours solely to administer a released TAKS, and 30 to 90 minutes for the KeyMath3.

Overall, the relationship between the time to take the STEEP3M and the score on the test was negative, indicating that as a student took more time, their score decreased.

Students who are more proficient in mathematics would evidence greater fluency in answering questions and spend less time thinking or making false starts, and therefore work more quickly. Additionally, as reported in Chapter II, a student's reading level will impact performance on mathematics tests; it takes a poor reader longer to complete a mathematics test. Since the average time to take the STEEP3M was 13.69 minutes, an administration time of 15 minutes or less would increase efficiency.

Summary of Results

Analysis of three quantitative data sources (STEEP3M, GRADE, and TAKS) and one qualitative data source (T-SIQ) were conducted to examine the usefulness of the STEEP3M as an assessment tool. Three elementary schools, 337 students, and 22 educators participated in the study. Overall, the STEEP3M was quite an easy test, with an average item difficulty IQR of .67 to .86. The Cronbach's Alpha for the STEEP3M was .88, indicating that the questions comprising the test are internally consistent. Criterion-related validity analysis of the STEEP3M showed that it is a reasonably good predictor within the lowest third of the class by GRADE.

Logistic regression and ROC analysis provided options for optimized solutions that could be used for grouping students for instruction and planning. Teachers indicated they would consider using the STEEP3M as a part of a decision-making rubric to identify struggling students; however, they would not use it as the sole criteria. The time involved to administer, score, and interpret the test makes the test more efficient to use than other measures. The relationship between a student's time to take the test and their score was negative, indicating that the longer the student spent taking the test, the more the score decreased. Descriptive statistic analyses were conducted on the STEEP3M sub-groupings, indicating poor discriminability. Instructional staff would not find this information useful for diagnostic decisions or lesson planning.

Correlation between the STEEP3M and the TEKS 3-1 was the only correlation in the acceptable range. The majority of the correlations between the STEEP3M item sub-groupings and TEKS objectives were weak, and will be further examined in Chapter V.

CHAPTER V

DISCUSSION AND SUMMARY

Purpose of the Study

The purpose this study was to examine the quality and utility of a universal screener assessment tool, the *System to Enhance Educational Performance Grade 3 Focal Mathematics Assessment Instrument (STEEP3M)*. The STEEP3M is a brief standards-based formative assessment of third grade mathematics. This study included three elementary schools for the one-time administration of the STEEP3M assessment to 337 third graders. It was given at or about the same time that they took the Grade 3 Texas state standards-based mathematics assessment, the Texas Assessment of Knowledge and Skills (TAKS), in the Spring 2007. The campus-based administration of the STEEP3M was followed by interviews with 22 instructional staff involved in third grade mathematics curriculum. The purpose of this study was to examine the following four qualities of the STEEP3M:

1. Internal consistency and criterion-related concurrent validity of the STEEP3M;
2. Screening students for a multi-tiered decision-making process such as for grouping or placement;
3. Utility for instructional planning and intervention recommendations; and
4. Efficiency of administration, scoring, and reporting results.

Although increasing student performance on mathematics state standards is a national priority and one of the major goals of the American educational system (Clarke & Shinn, 2004; Jordan, Kaplan, Locuniak, & Ramineni, 2007; National Educational Goals Panel, 2002), in 2003 fewer than 33% of fourth graders demonstrated proficient skills on the NAEP mathematics test (Manzo & Galley, 2003; VanDerHeyden & Burns, 2008). As reviewed in Chapters I and II, determining a formative CBM which is predictive of performance on high-stakes standards-based assessment is important to target students and identify content not mastered (Fletcher, 2005; VanDerHeyden &

Witt, 2005). The Texas Education Agency (TEA) has taken the position that the TAKS serves as a universal screener (TEA, 2009c) to identify students at-risk of failing standards-based assessment. In Texas, however, the mathematics TAKS test is not given until third grade. Therefore another screening instrument is needed to identify third graders at-risk of failure. According to its developers, an instrument for doing that is the STEEP3M, a brief mathematics curriculum based measurement (CBM) tool for predicting performance on the TAKS.

This study focused on universal screening of mathematics skills in the whole-group setting using the STEEP3M, an instrument reported to be aligned with curriculum that is national (National Council of Teachers of Mathematics [NCTM]) or state standards based (Texas Essentials Knowledge and Skills [TEKS]). This chapter is organized to provide summary information and discussion on method, research results by question, research limitations, and implications.

Conduct of the Study

This unfunded study was conducted between April and September 2007, on students enrolled in Grade 3 in Fall 2006. A convenience sample of three elementary campuses with third grade classrooms in three separate Local Education Agencies (LEAs) provided 337 students and 22 instructional staff for this study. Selection and profiles of the LEAs and participants are provided in Chapter III of this dissertation. Under the No Child Left Behind Act (2001), most students (97% to 99%) in the public school setting in the U.S. are required to take the regular enrolled grade level (EGL) standards-based assessment. In Texas this is the TAKS. Therefore, only students who were eligible to take the regular EGL TAKS were selected for the sample. Information on LEAs and participants was collected from the Texas Education Agency (TEA) Academic Excellence Indicator System (AEIS) and Lonestar public access data systems.

Using a standard set of instructions, the STEEP3M was administered one time to each of 18 classrooms with a total of 355 third graders, at three LEAs on three different days, one campus per day. The developers of the STEEP3M recommend a maximum of 15 minutes for the student to complete the assessment, but the time was extended to 20

minutes to allow for additional data collection for this correlational study. The researcher completed the administration, scoring, and data input.

Two indicators of student performance were selected as indicators of mastery of third grade mathematics curriculum: end-of-year third grade mathematics course grades (GRADE), and the scale score on the third grade mathematics TAKS. Information was gathered from instructional staff using a semi-structured interview instrument consisting of a predetermined set of open-ended questions, the *Teacher Survey Interview Questionnaire (T-SIQ)* developed by the researcher. The T-SIQ was administered to 22 instructional staff on a one-to-one basis rather than as a self-administered questionnaire in order to gather information on CBM screeners currently used, methodology for determining lesson planning, initial perception of the STEEP3M, and use of released TAKS tests. In all, five sources of data were needed to answer the four research questions for this study: a) AEIS, b) GRADE, c) STEEP3M, d) TAKS, and e) T-SIQ.

During the study, several unexpected events occurred which caused the researcher to deviate from the original plan. First, students in the final sample were required to have available data from all three sources, the one-time administration of the STEEP3M, GRADE, and TAKS. Eighteen students who took the STEEP3M were eliminated from the study because their data from all three sources were not available. This reduced the final sample size to 337 students. Second, specific data on qualification for special education and related services, Section 504, English as a second language, or other educational special programs were not made available to the researcher at the time of the study. When the researcher attempted to follow up with the three campuses to gather this additional information, two of the three campus administrators were no longer assigned to the campus. Therefore, no comparison could be made to determine whether the STEEP3M would provide information to identify these particular students as at-risk and therefore in need of interventions in mathematics.

The original plan for the sample was to identify three LEAs that were all traditional independent school districts. In fact, the sample consisted of two elementary schools from independent school districts and one charter school. Since in Texas charter

schools are public schools, several were considered for the study if they met the criteria for selection as outlined in Chapter III.

The STEEP3M was provided to the researcher in black and white format consisting of 3 pages and 41 problems. No specific information was provided to the researcher on curriculum alignment even though the developers purported the instrument to be aligned to the NCTM focal points. Although separate from the NCTM standards, according to TEA (2009c), the TEKS are also aligned to the NCTM. Upon review of the STEEP3M, the researcher noted several things that could impact student performance.

First, on most items, no specific directions for how to answer the question were provided, either on the stimulus directions read aloud by the researcher or written on the test itself. Second, no examples were given for any item. Third, although reading to the student is an allowable accommodation on the TAKS (TEA, 2007, 2009), it was not an allowable accommodation on the STEEP3M. As noted in Chapter III, a student's reading ability and comprehension may impact their performance on a mathematics test without this accommodation. Further, the vocabulary on the test would require students to have an understanding of specific terms such as "estimate," "standard form," and "pattern." Without this accommodation, a student's reading comprehension level rather than mathematics knowledge may be what is being measured. Additionally, items on the test were shaded and did not appear clearly in several instances. Rather than black and white, the test should be printed in colors with high contrast such as blue and yellow, in order for the student to see the differentiation of figures. Last, a problem with the instrument itself was that minimal area was provided on the STEEP3M for students to do "scratch" written work.

These events notwithstanding, the STEEP3M is one of only a few main providers of commercial CBM formative assessments in the area of mathematics. Therefore, the researcher decided to continue the study, primarily because information from the literature reviewed in Chapter II indicated there is a lack of research on mathematics universal screeners aligned to state standards-based curriculum that are predictive of performance on the high-stakes end-of-year test.

Research Results

Descriptive statistics were conducted on the STEEP3M, GRADE, and TAKS. All students scored at least 4 items correctly, with no students answering all 41 items correctly. The STEEP3M had a negatively skewed distribution, indicating that the test was too easy, with many high scores and few low scores (Huck, 2004; Sax, 1989). Only six students received a GRADE in the failing range, which impacted the negative skewing of the distribution. The teacher assigned GRADE is an aggregate of individual assessments that usually includes many subjective measures (Allen, 2005; Marzano, 2000). Although used as a measure of a subject's mastery and the basis for determining whether a student will progress to the next grade level, the GRADE has unknown, probably low, validity, which may make it hard to measure strong relationship with another variable. Nevertheless, it is a standard in education for measuring a student's mastery of grade level curriculum. The TAKS scale score distribution for the study was almost a perfect bell shape and considered normal. Performance by students in the study was similar to performance by students as a whole in the state on the third grade mathematics TAKS.

Research Question One

How well does the mathematics curriculum based measurement (CBM) assessment tool, the STEEP3M, reflect end-of-year student performance on state standards-based curriculum as measured by the end-of-year high-stakes TAKS test? (Internal consistency and criterion-related concurrent validity.)

Research Question One dealt with instrument qualities, primarily focusing on two psychometric features of the STEEP3M as measured by internal consistency and criterion-related concurrent validity. Four data sources were needed to answer this question: a) AEIS; b) GRADE (ordinal); c) STEEP3M (interval); and d) TAKS scale score (interval). Logistic regression using the STEEP3M scores was conducted to predict two dichotomous variables: GRADE (pass/fail) and TAKS (pass/fail). Classical item analysis was used to determine item difficulty, item discrimination, and the Cronbach's Alpha. Overall the STEEP3M was quite an easy test. The STEEP3M had acceptable

discrimination even though the items were not evenly distributed with mostly easy items and a few difficult ones. The questions comprising the STEEP3M were internally consistent with approximately 10% of the items needing revision.

The student sample was subdivided into three groups for analysis for criterion related concurrent validity. The STEEP3M predicted both the GRADE and TAKS moderately well. A follow-up multiple regression was conducted predicting STEEP3M from a combination of GRADE and TAKS, which resulted in a slightly better result than either of the independent relationships. This means that although they are separate independent measures and share considerable variance, they tend to measure the same thing. The student's mathematics ability and study habits probably are the common ingredient. Based on these analyses and the negative skewing of the distribution educators would only be able to use the results of the STEEP3M as a predictor of failing performance on the end-of-year TAKS test for the lowest third of the class. Results would not be useful for prediction for the middle or upper third of the class. As currently written the STEEP3M does not adequately reflect end-of-year performance for two-thirds of the students in the study.

Research Question Two

To what extent does the STEEP3M provide information to support a multi-tiered decision-making process in the third grade school setting, such as for grouping and placement of students?

Research Question Two focused on the utility of the STEEP3M for a multi-tiered decision-making process such as for grouping and placement of students. The STEEP3M was compared to the strongest two existing "true criteria" for students requiring more intensive help: GRADE and TAKS. Four data sources were needed to answer this question: a) GRADE (ordinal); b) STEEP3M (interval); c) TAKS scale score (interval); and d) TAKS results by objective (interval). The question was answered by logistic regression, using the STEEP3M scores to predict two dichotomous variables: GRADE (pass/fail) and TAKS (pass/fail).

The standard method for judging accuracy in predicting pass/fail is by the Receiver Operator Curve (ROC). Central to using the ROC is a cut-off level to separate students into two groups. The two groups in this study were students in danger of failing the TAKS and those who were not (Allen & Yen, 2002; Crocker & Algina, 1986, 2002). Failing either the TAKS or GRADE has potential serious consequences for students, including retention, placement in special programs, or targeted intervention. The ROC is the plot of the true positive rate (sensitivity) against the false positive rate (specificity) of the different cut-off points for a diagnostic test. Sensitivity is how well the test can pick out students who will actually fail, versus specificity, which is how well the test picks out students who in fact won't fail. A test can provide useful information to support a multi-tiered decision-making process in the third grade school setting, such as for grouping and placement of students, if an optimized solution can be used to select an appropriate cut-off point where classification accuracy, sensitivity, and specificity are maximized. As reported by Fuchs et al. (2007), most studies on mathematics screeners have not reported this type of decision utility on sensitivity or specificity. This study does report the classification accuracy, sensitivity, and specificity.

Several optimized solutions were generated from statistical analysis as described in Chapter IV. From these, two made sense, but neither demonstrated that the STEEP3M maximized classification accuracy, sensitivity, or specificity. Even after considering the limited screening capability of the STEEP3M, if one of the two reasonable solutions were chosen, an excessive number of students would be identified for intervention. Fuchs et al. (2007) recommend using the sum of true and false positives as the sample for identifying students for interventions. The first reasonable solution would have identified 95 students for intervention when in reality only 64 students actually failed the TAKS in the sample. This represents an over-identification of nearly 50% for tiered interventions.

The second reasonable solution would yield an even higher number (129 students) for intervention. At this solution, twice as many students would be placed in interventions as necessary, which translates to costly staffing allocations, more

instructional time allocated, and the additional expense of remediation materials.

Although the goal of a screener is to cast a wide net which possibly might include some students not in need of intervention, educators want to minimize this type of error. Not only would these students receive costly and time-consuming unnecessary interventions, they would be held back from making progress commensurate with their peers, potentially creating a problem where there was none. None of the solutions indicated that the STEEP3M did a good job of accurately classifying students with sufficient sensitivity or specificity for a multi-tiered decision-making process in the third grade school setting, such as grouping or placement of students, without making one of the two types of errors.

Analysis of instructional staff perception based on the T-SIQ interviews indicated that the majority of the teachers would consider using the STEEP3M as a beginning-year screener. None indicated that they would use it as the sole criteria for placing students in more intensive interventions. Instead they said that they might consider using it as a part of a rubric for decision-making, combined with the previous year's mathematics course grades and assessment scores. Therefore, even if the STEEP3M had acceptable classification accuracy, it is unlikely that educators would at this point accept the results as sufficient information to support a multi-tiered decision-making process.

Research Question Three

To what extent does the STEEP3M permit teachers to generate detailed mathematics lessons and unit planning for whole class and individual needs?

Research Question Three focused on the STEEP3M sub-skills and the extent to which they matched the standards-based mathematics curriculum to provide information to instructional staff for planning instruction and targeting interventions. Six sources of data and information were used to answer this question: a) GRADE (ordinal); b) TEKS (state standards curriculum); c) STEEP3M (interval); d) TAKS scale score (interval); e) TAKS results by objective (interval); and f) T-SIQ. Analyses to answer this question

were conducted on quantitative data (descriptive, regression, and correlation) and qualitative data.

Although the STEEP3M did have items that measured each of the six TEKS objectives represented on the TAKS, most (80%) of the items on the STEEP3M were aligned to two objectives measuring either use of numbers, operations, and quantitative reasoning or patterns and relationships. Based on correlational analysis, these items were not a reasonable measure of mathematics sub-skill weakness for these two objectives. Since the other four TEKS objectives represented on the STEEP3M had few items on the test, little diagnostic information was provided that would help teachers plan lessons. In addition, none of the questions on the STEEP3M measured the key TEKS objectives on time, temperature, or money. Even though the developers of the STEEP3M reported alignment to national and state standards, statistical analysis did not support this claim. In comparison, other measures such as the KeyMath3, even though longer to administer, would provide better information for instructional planning such as detailed information on mathematical concepts and skills. Overall, instructional staff would not find information from the administration of the STEEP3M in its' current form useful for diagnostics.

The majority of the correlations between the STEEP3M and TAKS were weak. The lack of correlation was probably due to a lack of internal consistency in STEEP3M or TAKS, lack of fit in content between the two tests, or the very different formats of the tests. The TAKS contains sufficient items for all six objective areas, whereas the STEEP3M does not. Likewise, the format of the tests differs in length and type of question. The TAKS is primarily comprised of multiple-choice questions. In all, information from the results of the STEEP3M would not be useful to teachers to generate detailed mathematics lessons and unit planning for whole class and individual needs.

During interviews, the majority of the instructional staff indicated that they would consider using information from the STEEP3M to plan instruction. In reality, however, this appears unlikely since most also responded that teams of teachers meet in

the summer or before school starts to review and agree on the scope and sequence of curriculum objectives to teach. Instructional staff who reported this process for developing lesson plans also reported that they follow this sequence and do not have much input in selecting instructional materials or in changing sequencing based on student need. According to teachers interviewed, this scope and sequence was based neither on data from assessment of student needs nor on the previous year's assessment scores. Instead it was based on alignment to covering the TEKS.

None of the instructional staff interviewed reported that there is a current universal screener for mathematics used from textbook, commercial, or other instructional materials. Instructional staff primarily reported that they assess the student's mathematics skills by verbally questioning or individually testing the student using classroom materials. These methods are time-consuming, taking away from classroom instruction during the crucial first weeks of the school year. All instructional staff reported that an assessment that provides predictive information would be beneficial for planning tiered interventions or grouping for improving student performance on the TAKS. In summary, however, even though there was agreement on the need and usefulness of a universal screener, even if the STEEP3M had provided useful information, it is unlikely to impact changes under the current system at these three elementary schools.

Research Question Four

How efficiently can the one-time, group-administered STEEP3M be administered to third grade students (as indicated by time to administer, score, report results, and provide feedback to teachers)?

Research Question Four examined efficiency of the STEEP3M as measured by a) the time and cost effectiveness of administering the universal screener; and b) the relationship of the student's test time to their score as a measure of efficiently determining student proficiency level. Three sources of data were needed to answer Research Question Four: a) STEEP3M (interval); b) TAKS (interval); and c) T-SIQ.

Five statistical measures were used to answer this question: a) descriptive statistics; b) anecdotal information; c) observations; d) time recording; and e) logistic regression.

The average amount of time to take the STEEP3M in the large group setting was less than 14 minutes. Scoring and reviewing data was also not very time consuming. In comparison to individually administered tests, the administration of the STEEP3M would be far less involved or time consuming than a released TAKS or standardized academic achievement instrument such as the KeyMath3. Overall, the STEEP3M can be efficiently administered and scored, making it less time consuming and therefore more cost effective than other measures. The positive findings for this question are offset by the assessment's lack of correlation to TEKS and classification accuracy. Even though the STEEP3M can be quickly administered, little useful information is gathered.

Research Summary

Mathematics proficiency is a national education goal and has received recent attention, which has led to the establishment of a National Mathematics Advisory Panel. Concern for improving student performance in mathematics has produced an increasing focus with little significant change in outcomes (Kelley, Hosp, & Howell, 2008). Effective evaluation is necessary to ensure that mathematics instruction and intervention are aligned to student needs. Additionally, instructional staff must have knowledge of their standards based curriculum and their school's instructional methodology in order to evaluate student proficiency. CBM is a quick, efficient, and cost-effective method for assessing students' academic skills. Determining whether the CBM is aligned with the curriculum and has reliability and validity is integral to identifying good measures for use in the educational setting. Limited studies on CBM aligned with state standards have been conducted, and even fewer have been done on the predictive ability for high-stakes standards-based assessment. Those studies that have been conducted have not systematically provided information on classification accuracy, sensitivity, and specificity. Therefore, the overall purpose of this study was to examine how well the STEEP3M reflected end-of-year student performance on state standards-based

curriculum and provided needed information on a content domain, and how practical it was to administer (Foegen, Jiban, & Deno, 2007).

An important contribution of this study to the broader research on mathematics CBM assessment instruments is the sample size, coverage of a content domain, and student grade level. Recent meta-analyses on studies on CBMs for mathematics found that most, like this study, focused on elementary mathematics assessments; however, only one of 41 studies was conducted on a sample that had more students than the 337 in this study (Foegen, Jiban, & Deno, 2007; Fuchs et al., 2007). Two groups of funded researchers, one headed by Clarke and one headed by VanDerHeyden, account for most of the published studies in early mathematics (Foegen, Jiban, & Deno, 2007). The studies by the Clarke group have focused on limited mathematics skills such as quantity discrimination or identifying the missing number. The studies by the VanDerHeyden group primarily have been conducted with prekindergarten and kindergarten students. Additionally, this study reported classification accuracy, sensitivity, and specificity.

In 1989, Shinn recommended that the typical administration procedures require administering three 1- to 3-minute CBM probes, and the mean be used to derive student performance. Recent guidelines have deviated from this recommendation, concluding that single- and multiple-skill CBM can be used (Christ, Scullin, Tolbize, & Jiban, 2008). Probes of 1 to 3 minutes might be sufficient to measure single skills; however, longer assessments are needed to assess multiple skills, as is the case with mathematics. As the length and duration of the assessment increases, so does its dependability and generalizability (Christ, Scullin, Tolbize, & Jiban, 2008). Probes of 1 to 4 minutes may be sufficient for low-stakes decisions, but assessments of approximately 14 minutes or longer are needed for high-stakes or criterion-referenced decisions (Christ, Scullin, Tolbize, & Jiban, 2008). Minimal research has been conducted on longer assessments such as the STEEP3M in the context of high-stakes, RtI, or criterion-referenced decisions. This study provides information to increase the body of research on longer formative assessments.

Although computational skills are essential components of mathematics, they are insufficient for success in the elementary classroom or on the high-stakes state standards-based test (Fuchs, Fuchs, & Zumeta, 2008). Most CBM measures developed for mathematics have focused on computational skills. Even though heavily weighted on two TEKS objectives, all six objectives as measured on TAKS, though some minimally, were represented on the STEEP3M. Developers of the STEEP3M have attempted to sample a content domain for third grade mathematics. Information from this study on the STEEP3M formative assessment, which was developed using the curriculum-sampling approach, provides needed information for educators and for the test developers on how to improve the instrument. No additional research on the STEEP3M in its current form would be necessary. Additional research might be necessary once the developers make recommended changes to the instrument.

Although various fluency measures have been identified in the area of reading, fewer studies have been conducted in the area of mathematics. The degree to which a student can respond rapidly and correctly to a particular set of mathematics problems reflects how well that student has mastered a set of skills for a content area. This information simultaneously informs instruction (VanDerHeyden & Burns, 2008). Criteria for reading fluency have been developed; however, fewer have been suggested for the area of mathematics. Information provided to answer the research questions on this study may prove useful towards developing mathematics fluency criteria. VanDerHeyden and Burns (2008) utilized information from the ROC to find fluency scores to match a specificity level for proficiency criteria. Future studies might be extended to include this review of the ROC if the developers were to utilize information from this study to make needed changes to the STEEP3M.

Research Limitations

Several limitations were noted during the conduct of the study noted as follows.

Design Limitations

Given the sample composition, size, and focus of the study, the results cannot be used to infer conclusions beyond the study's scope. The study was geographically

restricted using a convenience sample from three elementary schools in the southeast Texas area. Although the sample was diverse, the individual campuses were not necessarily so. The results, however, may expand on the body of research on what to research to examine the accuracy of a brief CBM as an indicator of performance on standards-based assessment, provide information regarding the validity for planning instruction, and evaluate the ability of such a CBM to aid in screening students for a multi-tiered decision-making process. Although future studies on the current form of the STEEP3M are not recommended, any future studies on instruments subsequently developed might include more geographic regions in Texas.

Timeline of Study

The study was conducted using concurrent validity, the STEEP3M was administered close to the time of the Spring 2007 third grade mathematics TAKS. Since the students had already received third grade classroom instruction for most of the school year, this may have impacted the results of their performance on the STEEP3M. As a formative assessment, the STEEP3M would have been administered to incoming third graders in Fall 2006. Again, although no future studies on this instrument are needed, future studies might be designed to administer the formative assessment would be administered in the fall, and then compare it to the actual performance on the TAKS administered the following spring. Designing a future study in this way would provide actual results rather than predicted ones.

Statewide Assessment Limitations

A potential limitation of the study would be if the Texas Education Agency or Texas policymakers made changes in the state standards (TEKS) or decided to utilize another assessment tool or method, such as end-of-course examination. The life expectancy for high-stakes standards-based assessment has been about 8 to 10 years (TEA, 2009). Changes such as use of end-of-course examination might be on the horizon for elementary schools. Future studies would need to review the TEKS to determine whether there were any changes. Since the STEEP3M reports to be aligned to the NCTM

and therefore to the TEKS, changes in the standards-based assessment would change to the extent to which different national or state standards-based objectives were measured.

Time Limit on Administration

According to TEA, most students complete the TAKS in 2 to 3 hours; however, it is an untimed test. Students who do not complete the test within that timeframe have the option for extended, unlimited time. A potential limitation of this study might be that the time limit on the screener's administration may not allow students sufficient time to demonstrate the full range of their knowledge and skills. Since the focus of the study is to identify students who demonstrate accuracy and fluency in completing the assessment, the design of the study centered on utilizing a screener that is efficient and easy to administer. Also offsetting this limitation is recent literature indicating that longer assessments (e.g., 14 minutes or longer) are needed to measure multiple skills for criterion-referenced or high-stakes decisions. Future studies might limit the administration time to 15 minutes or expand it to 25 or 30 minutes, to determine the impact this would have on the study.

Summary Conclusions

Math skills are essential to daily life, impacting a person's ability to function at home, work, and in the community. Although reading has been the focus in recent years, many students struggle in math. The inability to master math calculation and problem solving has contributed to the rising incidence of student failure, referrals for special education evaluations, and dropout rates. Studies have shown that curriculum based measurement (CBM) is a well-established tool for formative assessment, and could potentially be used for other purposes such as a prediction of state standards test scores, however to date there are limited validity studies between mathematics CBM and standard-based assessment. This research examined a brief assessment that reported to be aligned to national curriculum standards in order to predict student performance on state standards-based mathematics curriculum, identify students at-risk of failure, and plan instruction. Evidence was gathered on the *System to Enhance Educational Performance Grade 3 Focal Mathematics Assessment Instrument* (STEEP3M) as a

formative, universal screener. Using a sample of 337 students and 22 instructional staff, four qualities of the STEEP3M were examined: a) internal consistency and criterion related validity (concurrent); b) screening students for a multitiered decision-making process; c) utility for instructional planning and intervention recommendations; and d) efficiency of administration, scoring, and reporting results which were the basis of the four research questions for this study.

Overall the STEEP3M was a relatively easy test with about 10% of the test needing revision. As currently written the STEEP3M did not adequately reflect end-of-year performance on state standards-based curriculum as measured by the high stakes state assessment or end-of-course grades for two-thirds of the students in the study. Several optimized solutions were generated from Receiver Operator Curve (ROC) statistical analysis; however none demonstrated that the STEEP3M maximized classification accuracy, sensitivity, or specificity needed to support a multi-tiered decision-making process for third grade mathematics. Although the STEEP3M did have items that measured each of the six TEKS objectives represented on the TAKS, most (80%) of the items on the STEEP3M were aligned to only two objectives. Based on correlational analysis, unfortunately these items were not a reasonable measure of mathematics sub-skill weakness even for these two objectives. Since the other four TEKS objectives represented on the STEEP3M had few items on the test, little diagnostic information was provided that would help teachers plan lessons. In addition, none of the questions on the STEEP3M measured the key TEKS objectives on time, temperature, or money. Even though the developers of the STEEP3M reported alignment to national and state standards, statistical analysis did not support this claim.

In semi-structured interviews instructional staff reported that they currently there is no universal screener consistently used to identify students struggling in math prior to the end of Grade 3. Although instructional staff reported they would consider using the STEEP3M as a part of a decision-making rubric along with other measures, even if the STEEP3M provided useful information, it is unlikely they would do so. Responses to the

T-SIQ indicated that lessons are developed before the school year starts, more in response to the sequence of the state standards than to students' needs.

Although practical and efficient to administer, the test does not provide sufficient information on the content domain and does not accurately classify students in need of assistance. The average amount of time to take the test was just under 14 minutes, which recent literature indicates is an appropriate length for criterion-referenced and high-stakes decision-making. Future studies would not be recommended without changes to the instrument. In summary, the STEEP3M CBM is not an adequate one-point, static, formative assessment and does not appear to be best suited as an indicator of student performance on state standards.

NOMENCLATURE

| | |
|-------|---|
| AEIS | Academic Excellence Indicator System |
| AYP | Adequate Yearly Progress |
| CBA | Curriculum Based Assessment |
| CBM | Curriculum Based Measurement |
| CRA | Criterion-Referenced Assessment |
| CRT | Criterion-Referenced Test |
| EGL | Enrolled Grade Level |
| GRADE | Teacher Reported End-of-Year Classroom Grades |
| IDEA | Individuals with Disabilities Education Improvement Act of 2004 |
| IEL | Institute for Educational Leadership |
| IQR | Interquartile Range |
| LEA | Local Education Agency |
| LEA1 | Local Education Agency 1 of Study |
| LEA2 | Local Education Agency 2 of Study |
| LEA3 | Local Education Agency 3 of Study |
| MD | Mathematics Disability |
| NAEP | National Assessment of Educational Performance |
| NCLB | No Child Left Behind Act of 2001 |
| NCSS | Number Cruncher Statistical System |
| NCTM | National Council of Teachers of Mathematics |
| NMAP | National Mathematics Advisory Panel |
| NRP | National Reading Panel |
| RD | Reading Disability |

| | |
|---------|---|
| ROC | Receiver Operator Curve |
| RtI | Response to Intervention |
| SD | Standard Deviation |
| SEA | State Education Agency |
| SRT | Standards Reference Test |
| STEEP | System to Enhance Educational Performance |
| STEEP3M | System to Enhance Educational Performance Grade 3 Focal Mathematics Assessment Instrument |
| TABS | Texas Assessment of Basic Skills |
| TAKS | Texas Assessment of Knowledge and Skills |
| TAMU | Texas A&M University |
| TEA | Texas Education Agency |
| TEAMS | Texas Educational Assessment of Minimum Skills |
| TEC | Texas Education Code |
| TEKS | Texas Essential Knowledge and Skills |
| TLI | Texas Learning Index |
| T-SIQ | Teacher Survey Interview Questionnaire |
| USDOE | United States Department of Education |
| USDOL | United States Department of Labor |

REFERENCES

- Allen, J. (2005). Grades as valid measures of academic achievement of classroom learning. *The Clearing House*. Retrieved July 29, 2009, from <http://www.tcnj.edu/~senate/resources/documents/GradesasValidMeasures.pdf>
- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.
- Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory* (2nd ed.). Long Grove, IL: Waveland Press.
- Allington, R. (1994). What's special about special programs for children who find learning to read difficult? *Journal of Reading Behavior*, 26(1), 95-115.
- Anrig, G. R., & LaPointe, A. E. (1989). What we know about what students don't know. *Educational Leadership*, 47(3), 4-9.
- Ardoin, S., Sulto, S., Witt, J., Aldrich, S., & McDonald, E. (2005). Accuracy of readability estimates' predictions of CBM performance. *School Psychology Quarterly*, 20(1), 1-22.
- Ardoin, S., Witt, J., Sulto, S., Connell, J., Koenig, J., Resetar, J., Slider, N., & Williams, K. (2004). Examining the incremental benefits of administering a maze and three versus one curriculum-based measurement reading probes when conducting universal screening. *School Psychology Review*, 33(2), 218-233.
- Badian, N. (1983). Dyscalculia and nonverbal disorders of learning. In H. Myklebust (Ed.), *Progress in learning disabilities: Vol. 5* (pp. 235-264). New York: Grune & Stratton.
- Baker, S., Gersten, R., & Lee, D. (2002). A synthesis of empirical research on teaching of mathematics to low-achieving students. *The Elementary School Journal*, 103(1), 51-73.
- Berkas, N., & Pattison, C. (2007). Closing the achievement gap. *NCTM News Bulletin*, 43(9), 9.
- Blankenship, C. (1985). Using curriculum based assessment data to make instructional decisions. *Exceptional Children*, 52, 233-238.
- Boaler, J. (2009). When politics took the place of inquiry: A response to the National Mathematics Advisory Panel's review of instructional practices. *Educational Researcher*, 37(9), 588-594.

- Bottge, B., Rueda, E., LaRoque, P., Serlin, R., & Kwon, J. (2007). Integrating reform-oriented math instruction in special education settings. *Learning Disabilities Research & Practice*, 22(2), 96-109.
- Bryant, D. (2007). *Mathematics intervention for primary grades students at-risk for mathematics difficulties*. Retrieved May 15, 2007, from http://www.texasreading.org/utcrcla/pd/SERP_Math/downloads.asp
- Bryant, D., & Bryant, B. (2006a). *Emerging model: Three-tier mathematics assessment & intervention model*. Retrieved May 15, 2007, from http://www.texasreading.org/utcrcla/pd/SERP_Math/downloads.asp
- Bryant, D., & Bryant, B. (2006b). *Three-tier mathematics assessment & intervention model*. Retrieved May 15, 2007, from http://www.texasreading.org/utcrcla/pd/SERP_Math/downloads.asp
- Burrill, G. (1997). The NCTM Standards: Eight years later. *School Science and Mathematics*, 97, 335-339.
- Bus, A., & van IJzendoorn, M. (1999). Phonological awareness and early reading: A meta-analysis of experimental training studies. *Journal of Educational Psychology*, 91(3), 403-414.
- Busch, T., & Espin, C. (2003). Using curriculum-based measurement to prevent failure and assess learning in the content area. *Assessment for Effective Intervention*, 28(3&4), 49-58.
- Butler, S., & McMunn, N. (2006). *A teacher's guide to classroom assessment: Understanding and using assessment to improve student learning*. San Francisco: Josey.
- Calhoon, M. (2008). Curriculum based measurement for mathematics at the high school level: What we do not know...what we need to know. *Assessment for Effective Intervention*, 33(4), 234-239.
- Carnine, D., & Granzin, A. (2001). Setting learning expectations for student with disabilities. *School Psychology Review*, 30(4), 466-472.
- Carnine, D., Jitendra, A., & Silbert, J. (1997). A descriptive analysis of mathematics curricular materials from a pedagogical perspective: A case study of fractions. *Remedial and Special Education*, 18(2), 66-81.
- Carnine, D., Jones, E. D., & Dixon, R. (1994). Mathematics: Educational tools for diverse learners. *School Psychology Review*, 23, 406-427.

- Carpenter, R. (1985). Mathematics instruction in resource rooms: Instruction time and teacher competence. *Learning Disability Quarterly*, 8, 95-100.
- Carpenter, T. P., Matthews, W., Lindquist, M. M., & Silver, E. A. (1984). Achievement in mathematics: Results from the national assessment. *The Elementary School Journal*, 84, 485-495.
- Cawley, J. F., & Miller, J. H. (1989). Cross-sectional comparisons of the mathematical performance of children and learning disabilities: Are we on the right track toward comprehensive programming? *Journal of Learning Disabilities*, 22, 250-254, 259.
- Center for Public Education (CPE). (2007). *Standards-based tests*. Retrieved May 15, 2007, from http://www.centerforpubliceducation.org/site/c.kjJXJ5MPIwE/b.2506157/k.4CF9/Standardsbased_tests_illustration.htm
- Chambers, J., Cleveland, W., Kleiner, B., & Tukey, P. (1983). *Graphical Methods for Data Analysis*. Emeryville, CA: Wadsworth Publishing Co.
- Chard, D., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention*, 30(2), 3-14.
- Christ, T., & Vining, O. (2006). Curriculum-based measurement procedures to develop multiple-skill mathematics computation probes: Evaluation of random and stratified stimulus-set arrangements. *School Psychology Review*, 35(3), 387-400.
- Christ, T., Scullin, S., Tolbize, A., & Jiban, C. (2008). Implications of recent research: Curriculum based measurement of math computation. *Assessment for Effective Intervention*, 33(4), 198-205.
- Clarke, B., & Shinn, M. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review*, 33(2), 234-248.
- Cobb, P., & Jackson, K. (2009). The consequences of experimentalism in formulating recommendations for policy and practice in mathematics education. *Educational Researcher*, 37(9), 573-581.
- Colorado Department of Education. (2007). Response to intervention (RtI): Colorado system for student success-tiers of RtI. Retrieved June 23, 2007, from <http://www.cde.state.co.us/cdesped/RTI.asp>

- Confrey, J., Maloney, A., & Nguyen, K. (2009). Breaching the conditions for success for a National Math Advisory Panel. *Educational Researcher*, 37(9), 631-637.
- Connell, D. (2005). *Constructing math applications, curriculum-based assessment: An analysis of the relationship between applications problems, computation problems, and criterion-referenced assessments*. Unpublished doctoral dissertation, Louisiana State University, Lafayette.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.
- Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory* (4th ed.). Mason, OH: Cengage Learning.
- Cronbach, L. J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.
- Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on Coefficient Alpha and successor procedures. *Educational and Psychological Measurement*, 64(3), 391-418.
- Deno, S. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, 52, 219-232.
- Deno, S. (2003). Curriculum-based measures: Development and perspectives. *Assessment for Effective Intervention*, 28(3-4), 3-12.
- Deno, S., & Mirkin, P. (1977). *Data based program modification: A manual*. Reston, VA: Council for Exceptional Children.
- Dillman, D. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.
- Dirks, E., Spyer, G., vanLieschout, E., & Sonnevile, L. (2008). Prevalence of combined reading and arithmetic disabilities. *Journal of Learning Disabilities*, 41(5), 460-473.
- Duhachek, A., Coughlan, A. T., & Iacobucci, D. (2005). Results on the standard error of the Coefficient Alpha index of reliability. *Marketing Science*, (24)2, 294-301.
- Duhon, G., Noell, G., Witt, J., Freeland, J., Dufrene, B., & Gilbertson, D. (2004). Identifying academic skill and performance deficits: The experimental analysis of brief assessments of academic skills. *School Psychology Review*, 33(3), 429-443.

- Education Commission of the States (ECS). (2002). *No Child Left Behind issue brief: A guide to standards-based assessment*. Retrieved May 14, 2007, from <http://www.ecs.org/clearinghouse/35/50/3550.pdf>
- Elementary and Secondary Education Act (ESEA) of 1965. Pub. L. No. 89-10. 20 USC 70.
- Englert, C. S., Culatta, B. E., & Horn, D. G. (1987). Influence of irrelevant information in addition word problems on problem solving. *Learning Disability Quarterly*, 10, 29-36.
- Fletcher, J. (2005). Predicting math outcomes: Reading predictors and comorbidity. *Journal of Learning Disabilities*, 38(4), 308-312.
- Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Education*, 41(2), 121-139.
- Fuchs, L., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, 53, 199-208.
- Fuchs, L., & Fuchs, D. (1996). Connecting performance assessment and curriculum based measurement. *Learning Disabilities Research & Practice*, 11, 192-196.
- Fuchs, L. & Fuchs, D. (2007). A model for implementing responsiveness to intervention. *Teaching Exceptional Children*, 39(5), 14-20.
- Fuchs, L., Fuchs, D., Compton, D., Bryant, J., Hamlett, C., & Seethaler, P. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Exceptional Children*, 73(3), 311-330.
- Fuchs, L., Fuchs, D., & Courey, S. (2005). Curriculum-based measurement of mathematics competence: From computation to concepts and applications to real-life problem solving. *Assessment for Effective Intervention*, 30(2), 33-46.
- Fuchs, L., Fuchs, D., & Hamlett, C. (1989). Effects of instrumental use of curriculum-based measurement to enhance instructional programs. *Remedial and Special Education*, 10, 43-52.
- Fuchs, L., Fuchs, D., Hamlett, C., & Stecker, P. (1990). The role of skills analysis in curriculum-based measurement in math. *School Psychology Review*, 19, 6-22.
- Fuchs, L., Fuchs, D., & Hollenbeck, K. (2007). Extending responsiveness to intervention to mathematics at first and third grades. *Learning Disabilities Research & Practice*, 22(1), 13-24.

- Fuchs, L., Fuchs, D., Hosp, M., & Hamlett, C. (2003). The potential for diagnostic analysis within curriculum-based measurement. *Assessment for Effective Intervention*, 28(3-4), 13-22.
- Fuchs, L., Fuchs, D., & Zumeta, R. (2008). A curricular sampling approach to progress monitoring: Mathematics concepts and application. *Assessment for Effective Intervention*, 33(4), 225-233.
- Fuchs, L., Seethaler, P., Powell, S., Fuchs, D., Hamlett, C., & Fletcher, J. (2008). Effects of preventative tutoring on the mathematical problem solving of third-grade students with math and reading difficulties. *Exceptional Children*, 74(2), 155-173.
- Geary, D. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities*, 37, 4-15.
- George, G., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference, 11.0 update*. Boston: Allyn & Bacon.
- Gertsen, R., & Jordan, N. (2005). Early screening and intervention in mathematics difficulties: The need for action introduced to the special education series. *Journal of Learning Disabilities*, 38, 291-292.
- Ghiselli, E. (1963). Moderating effects and differential reliability and validity. *Journal of Applied Psychology*, 47, 81-86.
- Gickling, E., & Havertape, J. (1981). *Curriculum based assessment*. Minneapolis, MN: National School Psychology Inservice Training Network.
- Gickling, E., & Thompson, V. (1985). A personal view of curriculum based assessment. *Exceptional Children*, 52, 205-218.
- Gliem, J., & Gliem, R. (2003, October). Calculating interpreting, and reporting Cronbach's Alpha reliability coefficient for Likert type scales. In *2003 Midwest Research to Practice Conference in Adult, Continuing, and Community Education*, Columbus, OH.
- Goldenberg, C. (1994). Promoting early literacy development among Spanish speaking children: Lessons from two studies. In E. Hiebert & B. Taylor (Eds.), *Getting reading right from the start* (pp. 171-199). Boston: Allyn & Bacon.
- Good, R., & Jefferson, G. (1998). Contemporary perspectives on curriculum-based measurement validity. In M. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 61-68). New York: Guilford Press.

- Good, R., & Kaminski, R. (1996). Assessment for instructional decisions: Toward a proactive/prevention model of decision-making for early literacy skills. *School Psychology Quarterly, 11*, 1-11.
- Greeno, J., & Collins, J. (2008). Commentary on the final report of the National Mathematics Advisory Panel. *Educational Researcher, 37*(9), 618-623.
- Gross-Tsur, V., Manor, O., & Shaley, R. (1996). Developmental dyscalculia: Prevalence and demographic features. *Developmental Medicine and Child Neurology, 37*, 906-914.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
- Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale, NJ: Erlbaum.
- Hargis, C. (1987). *Curriculum based assessment: A primer*. Springfield, IL: Charles C. Thomas.
- Harris, C., Miller, S., & Mercer, C. (1995). Teaching initial multiplication skills to students with disabilities in general education classrooms. *Learning Disabilities Research & Practice, 10*, 180-195.
- Helwig, R., Anderson, L., & Tindal, G. (2002). Using a concept-grounded, curriculum-based measurement in mathematics to predict statewide test scores for middle school students with LD. *The Journal of Special Education, 36*(2), 102-112.
- Hiebert, E., & Taylor, B. (Eds.). (1994). *Getting reading right from the start*. Boston: Allyn & Bacon.
- Hintze, J. (2004). *NCSS 2004 and PASS 2002*. (Version 2004) [Computer software]. Kaysville, UT: Number Cruncher Statistical Systems.
- Hosmer, D., & Lemeshow, S. (2000) *Applied logistic regression* (2nd ed.). Hoboken, NJ: Wiley.
- Hosp, M., & Hosp, J. (2003). Curriculum-based measurement for reading, spelling, and math: How to do it and why. *Preventing School Failure, 48*(1), 10-17.
- Hosp, M., Hosp, J., & Howell, K. (2007). *The ABCs of CBM: A practical guide to curriculum-based measurement*. New York: Guilford.
- Huck, S. (2004). *Reading statistics and research* (4th ed.). Boston: Pearson.
- Huck, S. (2008). *Reading statistics and research* (5th ed.). Boston: Pearson.

- Idol, L., Nevin, A., & Paolucci-Whitcomb, P. (1999). *Models of curriculum-based assessment: A blueprint for learning*. Austin, TX: Pro-Ed.
- Idol-Maestas, L. (1983). *Special educators consultation handbook*. Rockville, MD: Aspen.
- Individuals with Disabilities Education Improvement Act (IDEA) of 2004, Pub. L. No. 108-446, 20 U.S.C. §§ 1400-1485, 34 C.F.R. pts. 300, 301.
- Institute for Educational Leadership (IEL): National Collaborative on Workforce and Disability. (2003). Math, science, and technology: Essential skills for career success in the 21st century. *Info Brief 7*. Retrieved June 23, 2007, from http://www.ncwd-youth.info/resources_&_Publications/information_Briefs/issue7.html
- Jitendra, A. (2005). Mathematics assessment: Introduction to the special issue. *Assessment for Effective Intervention*, 30(2), 1-2.
- Jitendra, A., Sczesniak, E., & Deatline-Buchman, A. (2005). An exploratory validation of curriculum-based mathematical word problem-solving tasks as indicators of mathematics proficiency for third graders. *School Psychology Review*, 34(3), 358-371.
- Jordan, N., Kaplan, D., Locuniak, M., & Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research & Practice*, 22(1), 36-46.
- Jorgenson, D. (1970). Prediction of predictability in the multivariate case. *Journal of Educational Measurement*, 7(3), 175-185.
- Kelly, A. (2009). Reflections on the National Mathematics Advisory Panel final report. *Educational Researcher*, 37(9), 561-564.
- Kelly, B., Hosp, J., & Howell, K. (2008). Curriculum based evaluation and math: An overview. *Assessment for Effective Intervention*, 33(4), 250-256.
- Ketterlin-Geller, L., McCoy, J., Twyman, T., & Tindal, G. (2003). How do critical thinking measures fit within standards-based reform? *Assessment for Effective Intervention*, 28(3&4), 37-48.
- King-Sears, M. (1994). *Curriculum-based assessment in special education*. San Diego, CA: Singular Publishing Group, Inc.

- Komatsu, C. (2004). *The importance of fluent component skills in mathematical comprehension*. Unpublished masters thesis, Louisiana State University, Lafayette.
- Kunsch, C., Jitendra, A., & Sood, S. (2007). The effects of peer-mediated instruction in mathematics for students with learning problems: A research synthesis. *Learning Disabilities Research & Practice*, 22(1), 1-12.
- Lembke, E., & Foegen, A. (2009). Identifying early numeracy indicators for kindergarten and first grade students. *Learning Disabilities Research & Practice*, 24(1), 12-20.
- Lembke, E., Deno, S., & Hall, K. (2003). Identifying an indicator of growth in early writing proficiency for elementary school students. *Assessment for Effective Intervention*, 28(3&4), 23-35.
- Lembke, E., Foegen, A., Whittaker, T., & Hampton, D. (2008). Establishing technically adequate measures of progress in early numeracy. *Assessment for Effective Intervention*, 33(4), 206-214.
- Lewis, C., Hitch, G., & Walker, P. (1994). The prevalence of specific arithmetic difficulties and specific reading difficulties in 9- to 10- year old boys and girls. *Journal of Child Psychology and Psychiatry*, 35, 283-292.
- Lindlof, T., & Taylor, B. (2002). *Qualitative communication research methods*. Thousand Oaks, CA: Sage Publications.
- Lobato, J. (2009). On learning processes and the National Mathematics Advisory Panel report. *Educational Researcher*, 37(9), 595-601.
- Maccini, P., Mulcahy, C., & Wilson, M. (2007). A follow-up of mathematics interventions for secondary students with learning disabilities. *Learning Disabilities Research & Practice*, 22(1), 58-74.
- Malmgren, K., McLaughlin, M., & Nolet, V. (2005). Accounting for the performance of students with disabilities on statewide assessments. *The Journal of Special Education*, 39(2), 86-96.
- Manzo, K., & Galley, M. (2003). Math climbs, reading flat on '03 NAEP. *Education Week*, (23)1, 81.
- Martson, D. (1989). A curriculum-based measurement approach to assessing academic performance: What is it and why do it. In M. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guilford.

- Martson, D., Deno, S., & Kim, D. (1995). Comparison of reading intervention approaches for students with mild disabilities. *Exceptional Children*, 62, 20-37.
- Marzano, R. (2000). *Transforming classroom grading*. Baltimore: Association for Supervision & Curriculum.
- Mastropieri, M., Bakken, J., & Scruggs, T. (1991). Mathematics instruction for individuals with mental retardation: A perspective and research synthesis. *Education and Training in Mental Retardation*, 26, 115-129.
- Merriam, S. (2001). *Qualitative research and case study applications in education*. San Francisco: Josey-Bass.
- Methe, S., Hintze, J., & Floyd, R. (2008). Validation and decision accuracy of early numeracy skill indicators. *School Psychology Review*, 37(3), 359-373.
- Microsoft. (2009). *Microsoft Excel homepage*. Retrieved March 26, 2009, from <http://office.microsoft.com/en-us/excel/default.aspx?ofcresset=1>
- Miller, S., & Hudson, P. (2007). Using evidence-based practices to build mathematics competence related to conceptual, procedural, and declarative knowledge. *Learning Disabilities Research & Practice*, 22(1), 47-57.
- Montague, M. (2007). Self-regulation and mathematics instruction. *Learning Disabilities Research & Practice*, 22(1), 75-83.
- Murphy, M., & Mazzocco, M. (2008). Mathematics learning disabilities in girls with Fragile X or Turner Syndrome. *Journal of Learning Disabilities*, 41(1), 29-46.
- National Assessment of Educational Performance (NAEP). (1992). *NAEP 1992 mathematics report card for the nation and the states* (Report No. 23-ST02). Washington, DC: National Center for Educational Statistics.
- National Assessment of Educational Performance (NAEP). (1996). *NAEP 1996 mathematics report card for the nation and the states* (NCES 97-488). Washington, DC: National Center for Educational Statistics.
- National Center for Education Statistics (NCES), U.S. Department of Education. (2006). Children 3 to 21 years old served in federally supported programs for the disabled, by type of disability: Selected years, 1976-77 through 2003-04. *Digest of Education Statistics, 2005* (NCES 2006-030) (chap. 2). Retrieved April 11, 2009, from <http://nces.ed.gov/fastfacts/display.asp?id=64>

- National Center on Student Progress Monitoring (NCSPM). (2007). *Frequently asked questions about curriculum-based measurement*. Retrieved May 27, 2007, from http://www.studentprogress.org/progressmonitoring_math_faq.asp#g14
- National Council of Teachers of Mathematics (NCTM). (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.
- National Council of Teachers of Mathematics (NCTM). (2007a). *Principles and standards for school mathematics*. Retrieved April 29, 2007, from <http://www.nctm.org/standards/default.aspx?id=58>
- National Council of Teachers of Mathematics (NCTMa). (2007b). *Research briefs*. Retrieved June 23, 2007, from <http://www.nctm.org/news/content.aspx?id=8468>
- National Educational Goals Panel. (2002). *Raising achievement and reducing gaps: reporting progress toward goals for academic achievement in mathematics*. Retrieved May 15, 2007, from <http://govinfo.library.unt.edu/negp/page5.htm>
- National Institute of Child Health and Human Development (NICHD). (2000). *Report of the National Reading Panel. Teaching children to read: an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups* (NIH Publication No. 00-4754). Washington, DC: U.S. Government Printing Office.
- National Mathematics Advisory Panel (NMAP). (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.
- National Reading Panel (NRP). (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institutes of Child Health and Human Development.
- National Research Council (NRC). (2006). The state of school mathematics in the United States. In J. Kilpatrick, J. Swafford, & B. Findell (Eds.), *Adding it up: Helping children learn mathematics*. Washington, DC: National Academy Press.
- NCLB & IDEA Rule, 34 C.F.R § 200, 300. (2007). Amendments to 34 C.F.R. parts 200 and 300: Title I – Improving the Academic Achievement of the Disadvantaged, No Child Left Behind (NCLB) Act; Individuals with Disabilities Education Improvement Act (IDEA): Final Rule. *Federal Register* 72, No. 67 (09 April, 2007).

- New Mexico Public Education Department. (2006). *Response to intervention: A systematic process to increase learning outcomes for all students – guidance for New Mexico schools*. Retrieved June 23, 2007, from www.ped.state.nm.us
- No Child Left Behind Act (NCLB) of 2001. Pub. L. No. 107-110, 20 USC 6301.
- Number Cruncher Statistical System (NCSS). (2009). *NCSS quick start and user manual*. Retrieved March 26, 2009, from <http://www.ncss.com/>
- O'Neill, K., & Stansbury, K. (2000). *Developing a standards-based assessment system: A handbook*. San Francisco: WestEd.
- Otaiba, S., & Fuchs, D. (2006). Who are the young children for whom best practices in reading are ineffective? An experimental and longitudinal study. *Journal of Learning Disabilities*, 39(5), 414-431.
- Oxford Dictionaries. (2007). *Mathematics*. Retrieved October 12, 2007, from http://www.askoxford.com/concise_oed/mathematics?view=uk
- Parmar, R., Cawley, J., & Frazita, R. (1996). Word problem solving by students with and without mild disabilities. *Exceptional Children*, 62(5), 415-429.
- Payne, D. A. (2003). *Applied educational assessment* (2nd ed.). Belmont, CA: Wadsworth.
- Pearson Assessments. (2009). *KeyMath3 Diagnostic Assessment publication summary form*. Retrieved May 26, 2009, from <http://www.pearsonassessments.com/keymath.aspx>
- Pikulski, J. (1997). Preventing reading problems: Factors common to successful early intervention programs. Retrieved June 23, 2007, from <http://www.eduplace.com/rdg/res/prevent.html>
- Putnam, D. (1989). *The criterion-related validity of CBM measures of math*. Unpublished master's thesis, University of Oregon, Eugene.
- Ralph, J. (1994). How effective are American schools? *Phi Delta Kappan*, 76(2), 144-152.
- Reese, C., Miller, K., Mazzeo, J., & Dossey, J. (1997). *NAEP 1996 report card for the nation and the states*. Washington, DC: National for Education Statistics.
- Reynolds, A. (1991). Early schooling of children at-risk. *American Educational Research Journal*, 28, 392-422.

- Riley, R. (1997). *Mathematics equals opportunity*. Washington, DC: Federal Department of Education. (ERIC Document Reproduction Service No. ED415119)
- Salkind, N. (2004). *Statistics for people who think they hate statistics*. Thousand Oaks, CA: Sage Publications.
- Savage, R., & Carless, S. (2004). Predicting curriculum and test performance at age 7 years from pupil background, baseline skills and phonological awareness at age 5. *British Journal of Educational Psychology*, 74, 155-171.
- Sax, G. (1989). *Principles of educational and psychological measurement and evaluation* (3rd ed.). Belmont, CA: Wadsworth.
- Schumacker, R. (2005). *Classical item analysis*. Retrieved May 28, 2009, from <http://www.appliedmeasurementassociates.com/White%20Papers/CLASSICAL%20TEST%20ANALYSIS.pdf>
- Shapiro, E. (1996). *Academic skills problems: Direct assessment and intervention* (2nd ed.). New York: Guilford.
- Shapiro, E., Edwards, L., & Zigmond, N. (2005). Progress monitoring of mathematics among students with learning disabilities. *Assessment for Effective Intervention*, 30(2), 15-32.
- Shepard, L. (2009). Commentary on the National Mathematics Advisory Panel recommendations on assessment. *Educational Researcher*, 37(9), 602-609.
- Shinn, M. (Ed.). (1989). *Curriculum-based measurement: Assessing special children*. New York: Guilford.
- Shippen, M., Houchins, D., Calhoun, M., Furlow, C., & Sartor, D. (2006). The effects of comprehensive school reform models in reading for urban middle school students with disabilities. *Remedial and Special Education*, 27(6), 322-328.
- Simpson, R., LaCava, P., & Graner, P. (2004). The No Child Left Behind Act: Challenges and implications for educators. *Intervention in School and Clinic*, 40(2), 67-75.
- Skiba, R., Magnusson, D., Martson, D., & Erikson, K. (1986). *The assessment of mathematics performance in special education: Achievement tests, proficiency tests, or formative evaluation?* Minneapolis, MN: Special Services, Minneapolis Public Schools.
- Snow, C., Burnes, M., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

- System to Enhance Educational Performance (STEEP). (2007a). *System to Enhance Educational Performance Grade 3 Focal Mathematics Assessment Instrument (STEEP3M)*. Unpublished copyrighted universal screener. Miami, FL: iSTEEP.
- System to Enhance Educational Performance (STEEP). (2007b). *System to Enhance Educational Performance (STEEP): About us*. Retrieved June 22, 2007, from http://www.isteep.com/about_us.html
- Stevenson, H. W., Chuansheng, C., & Lee, S. (1993). Mathematics achievement of Chinese, Japanese, and American children: Ten years later. *Science*, 257, 53-58.
- Stevenson, H. W., Lee, S.-Y., Chen, C., Stigler, J. W., Hsu, C.-C., & Kitamura, S. (1990). Contexts of achievement: A study of American, Chinese, and Japanese children. In H. W. Stevenson & J. W. Stigler (Eds.), *The learning gap*. New York: Summit.
- Stigler, J. W., Lee, S.-Y., & Stevenson, H. W. (1990). *Mathematical knowledge of Japanese, Chinese, and American elementary school children*. Reston, VA: NCTM.
- Sunderman, G., Orfield, G., & Kim, J. (2006). The principals denied by NCLB are central to visionary reform under NCLB. *The Education Digest*, 72(2), 19-24.
- Texas Education Agency (TEA). (2009a). *Academic Excellence Information System*. Retrieved May 15, 2009, from <http://ritter.tea.state.tx.us/perfreport/aeis/>
- Texas Education Agency (TEA). (2009b). *LONESTAR Education Reports*. Retrieved April 12, 2009, from <http://198.214.97.212/>
- Texas Education Agency (TEA). (2009c). *Texas Assessment of Knowledge and Skills: Assessment resources for teachers and administrators*. Retrieved April 12, 2009, from http://www.tea.state.tx.us/index3.aspx?id=3633&menu_id3=793
- Texas Essential Knowledge and Skills (TEKS) for Mathematics: Grade 3, Tex. Admin. Code (TAC) Title 19, part II §111.15 (1998 & Amended 2006).
- Thompson, P. (2009). On professional judgment and the National Mathematics Advisory Panel report. *Educational Researcher*, 37(9), 582-587.
- Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education* (6th ed.). Upper Saddle River, NJ: Merrill/Prentice-Hall.
- Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Merrill/Prentice-Hall.

- Thurber, R., Shinn, M., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review*, 31, 363-382.
- Tindal, G. & Parker, R. (1991). Identifying measures for evaluating written expression. *Learning Disabilities Research & Practice*, 6(4), 211-218.
- Tomlinson, C., Callahan, C., Tomchin, D., Eiss, N., Imbeau, M., & Landrum, M. (1997). Becoming architects of communities of learning: Addressing academic diversity in contemporary classrooms. *Exceptional Children*, 63, 269-282.
- U.S. Department of Education (USDOE). (1983). *National Commission on Excellence in Education: A nation at-risk*. Retrieved October 11, 2007, from <http://www.ed.gov/pubs/NatAtRisk/risk.html>
- U.S. Department of Education (USDOE). (2002). *Comprehensive school reform program*. Retrieved April 29, 2007, from <http://www.ed.gov/programs/compreform/2pager.html>
- U.S. Department of Education (USDOE). (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. (NCEE: 2004-3000). Washington, DC: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.
- U.S. Department of Education (USDOE). (2005). *Education in the United States: A brief overview*. Washington, DC: International Affairs Staff.
- U.S. Department of Education (USDOE). (2006a). *Mathematics framework for the National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board.
- U.S. Department of Education (USDOE). (2006b). *Secretary Spellings announces National Mathematics Advisory Panel members*. (Press release, May 15, 2006). Washington, DC: Author.
- U.S. Department of Labor (USDOL), Bureau of Labor Statistics. (1997). *Occupational outlook handbook*. Washington, DC: U.S. Government Printing Office.
- U.S. Department of Labor (USDOL), Bureau of Labor Statistics. (2007). *Occupational outlook handbook (OOH), 2006-07 Edition*. Retrieved June 23, 2007, from <http://www.bls.gov/oco/>
- VanDerHeyden, A., & Burns, M. (2008). Examination of the utility of various measures of mathematics proficiency. *Assessment for Effective Intervention*, 33(4), 215-224.

- VanDerHeyden, A., & Witt, J. (2005). Quantifying context in assessment: Capturing the effect of base rates on teacher referral and a problem-solving model of identification. *School Psychology Review*, 34(2), 161-183.
- VanDerHeyden, A., Witt, J., & Naquin, G. (2003). Development and validation of a process for screening referrals to special education. *School Psychology Review*, 32(2), 204-227.
- VanDerHeyden, A., Witt, J., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students. *School Psychology Review*, 30(3), 363-382.
- Vinchur, A. (1993). The prediction of predictability revisited: Differential predictability applied to managerial selection. *Educational and Psychological Measurement*, 53(4), 1085-1094.
- Washington Department of Education. (2007). *Special education: Response to intervention, universal screening, progress monitoring, and model programs, policies and procedures*. Retrieved June 23, 2007, from <http://www.k12.wa.us/SpecialEd/RTI.aspx>
- West Virginia Department of Education, Office of Special Education Achievement. (2007). *Response to intervention*. Retrieved June 23, 2007, from <http://wvde.state.wv.us/ose/RtI.html>
- Xin, Y., Jitendra, A., & Deatline-Buchman, A. (2005). Effects of mathematical word problem-solving instruction on middle school students with learning problems. *The Journal of Special Education*, 39(3), 181-192.
- Ysseldyke, J., and Algozzine, B. (2006). *Assessment for students with special needs*. Thousands Oaks, CA: Corwin Press.
- Zentall, S. S. (1990). Fact-retrieval automatization and math problem solving by learning disabled, attention-disordered, and normal adolescents. *Journal of Educational Psychology*, 82(4), 856-865.
- Zentall, S. S., & Ferkis, M. A. (1993). Mathematical problem solving for youth with ADHD, with and without learning disabilities. *Learning Disability Quarterly*, 16, 6-18.

APPENDIX A

TEACHER SURVEY INTERVIEW QUESTIONNAIRE (T-SIQ)

1. How does your campus assess the mathematics skills of students?
2. How do teachers make decisions on teaching mathematics?
3. What textbook is used for teaching mathematics on your campus?
4. Does your campus use commercial mathematics materials or programs?
 - a. If yes
 - i. Which one(s)?
 - ii. How are they used?
5. What is your initial perception of the focal point mathematics curriculum based assessment (CBA)?
 - a. Would you consider using this CBA?
 - i. If yes, how would you use this CBA?
 - b. In your opinion, would data from this CBA help you plan instruction?
 - i. If yes, how would you use data?
6. How long does it take to administer the state standards tests (Texas Assessment of Knowledge and Skills – TAKS) in the area of mathematics?
7. Do you use released TAKS with students?
 - a. If yes
 - i. How are they used?
 - ii. How long does it take to administer released TAKS?

VITA

Linda De Zell Hall received her Bachelor of Business Administration with majors in marketing and management from Texas Tech University, Lubbock, and received her Master of Education in educational psychology from Texas A&M University, College Station. She graduated with her Doctor of Philosophy from Texas A&M University, College Station, in December 2009 after completing the Special Education Leadership Interagency Collaboration (SELIC) program in educational psychology and education administration and human resources. As a part of the SELIC program, she demonstrated competencies in research, grant-writing, supervision, and teaching as well as providing service for educational administration leadership and principalship. She was a research team member for the Special Education Reading Interface Project, Master Reading Teacher Evaluation Project, Houston Area Initiative Research and Program Evaluation, Special Education Recruitment and Retention Grant Evaluation, and Institute for School University Partnerships Program Evaluation and Research.

Dr. Hall is employed at Region 4 Education Service Center (ESC) in Houston and serves on the Texas Education Agency Statewide Leadership Evaluation Network. She has served as a Director on the Council for Exceptional Children - Council for Educational Diagnostic Services (CEC-CEDS) National Board for Educational Diagnostician Certification. She is a Nationally Certified and a Texas Registered Professional Educational Diagnostician. Dr. Hall holds Texas State Board of Education certification as a Principal, Educational Diagnostician, and Teacher. Dr. Hall frequently presents at the state and national level on educating diverse learners; educational coaching and consultation; multidisciplinary team approach for evaluation; functional educational evaluation; school budgeting and finance; leadership and organizational capacity building; meeting facilitation; and legal mandates.

Dr. Hall can be reached at: 7145 West Tidwell, Houston, Texas. Her email is lhall@esc4.net.